

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平7-152499

(43) 公開日 平成7年(1995)6月16日

(51) Int.Cl.<sup>8</sup>

G 0 6 F 3/06

識別記号

5 4 0

3 0 5 C

庁内整理番号

F I

技術表示箇所

審査請求 未請求 請求項の数44 F D (全 31 頁)

(21) 出願番号 特願平6-209197

(22) 出願日 平成6年(1994)8月10日

(31) 優先権主張番号 特願平5-273200

(32) 優先日 平5(1993)10月5日

(33) 優先権主張国 日本 (J P)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 角田 仁

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(74) 代理人 弁理士 笹岡 茂 (外1名)

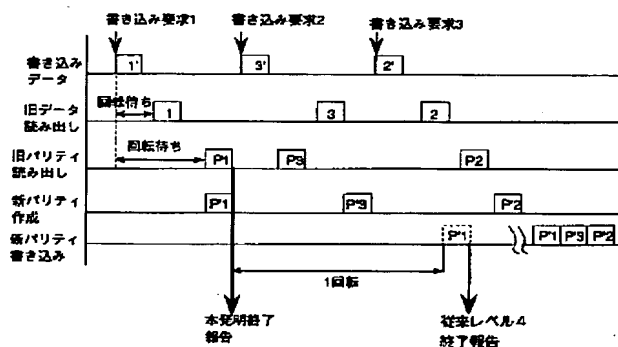
(54) 【発明の名称】 ディスクアレイ装置におけるパリティ格納方法、障害回復方法およびディスクアレイ装置

(57) 【要約】

【目的】 RAID (レベル4、5) のディスクアレイシステムにおいて、データの書き込みによるパリティの更新時のオーバーヘッドを減少させる。

【構成】 書き込み要求1、3、2が上位装置から送られ、書き込みデータ1'、3'、2'が送られると、旧データ1、3、2および旧パリティP1、P2、P3を夫々回転待ちの後読み出し、書き込みデータと旧データと旧パリティにより新パリティをP'1、P'2、P'3を作成し、キャッシュメモリに貯蔵しておき、新パリティが予めユーザが設定した値以上になった場合か、読み出し書き込み要求の発行されていないタイミングが生じた場合にパリティ格納用のドライブにまとめて書き込む。上位装置への書き込み処理終了報告は、新パリティをキャッシュメモリに格納した時点で行なわれる。パリティ格納用のドライブに替えてパリティ格納用のフラッシュメモリを用いることもできる。

図5



## 【特許請求の範囲】

【請求項 1】 上位装置に接続され、キャッシュメモリと少なくとも 1 台の制御装置を備えるディスクアレイコントローラと、該ディスクアレイコントローラに接続され複数台のディスク装置からなる少なくとも 1 つの論理グループを備えるディスクアレイユニットとを備え、前記ディスクアレイコントローラの制御装置が、前記上位装置から 1 回に読み出したまたは書き込みする単位で転送されてきたデータを分割せずに前記複数台のディスク装置の内の複数台のデータ用のディスク装置のいずれかに格納し、前記複数台のデータ用のディスク装置に格納されている各データによりパリティを生成し、この生成したパリティを、前記複数台のディスク装置の内のパリティ用のディスク装置に格納するディスクアレイ装置におけるパリティ格納方法であって、

このディスクアレイ装置に対し上位装置から発行された書き込み要求に対し、前記制御装置はパリティを更新した後、この更新したパリティを前記キャッシュメモリに格納し、同様に上位装置から発行されてきた別の書き込み要求に対し更新したパリティも該キャッシュメモリに格納し、これらの更新されたパリティを該キャッシュメモリ内に貯蔵し、このキャッシュメモリ内の更新されたパリティ群を書き込み要求の発行順に前記パリティ用のディスク装置の更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項 2】 請求項 1 記載のディスクアレイ装置におけるパリティ格納方法において、前記ディスクアレイコントローラに更新されたパリティを格納する専用メモリを設け、更新されたパリティを前記キャッシュメモリに替えて前記専用メモリに格納するようにしたことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項 3】 請求項 2 記載のディスクアレイ装置におけるパリティ格納方法において、前記専用メモリを揮発メモリとしたことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項 4】 請求項 1 記載のディスクアレイ装置におけるパリティ格納方法において、

前記キャッシュメモリにアドレス変換用テーブルを設け、

該テーブルには、パリティグループの各データの論理アドレスと該論理アドレスに対応するディスク装置番号およびディスク装置内アドレスと、パリティの論理アドレスと該論理アドレスに対応するディスク装置番号およびディスク装置内アドレスとパリティをキャッシュメモリ内に貯蔵した場合のキャッシュアドレスとキャッシュメモリ内に貯蔵されたパリティとディスク装置内に格納されているパリティの何れが有効かを示すフラグを備えることを特徴とするディスクアレイ装置におけるパリティ

格納方法。

【請求項 5】 請求項 4 記載のディスクアレイ装置におけるパリティ格納方法において、

更新されたパリティ群をシーケンシャルにまとめてディスク装置に書き込む途中において、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれているディスク装置のアドレスにおいては、一旦書き込みを中断し、書き込み要求が発行され、すでに更新されているディスク装置の無効なパリティのアドレスにおいて、シーケンシャル書き込みを再開することを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項 6】 請求項 4 記載のディスクアレイ装置におけるパリティ格納方法において、

更新されたパリティ群をシーケンシャルにまとめてディスク装置に書き込む際、ディスク装置の書き込み先の領域に、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれている場合、この有効なパリティをキャッシュメモリに読み出し、更新されたパリティと合わせてパリティ群を構成すると共にパリティの書き込み順を決定し、該パリティ群を該書き込み順にディスク装置の書き込み前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項 7】 請求項 1 記載のディスクアレイ装置におけるパリティ格納方法において、

前記パリティ用のディスク装置を複数台設け、更新されたパリティ群をまとめてディスク装置に書き込む際、複数のパリティ用のディスク装置に、更新されたパリティ単位に、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項 8】 請求項 1 記載のディスクアレイ装置におけるパリティ格納方法において、

前記パリティ用のディスク装置を複数台設け、更新されたパリティ群をまとめてディスク装置に書き込む際、複数のパリティ用のディスク装置に、バイト単位に、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項 9】 請求項 1 記載のディスクアレイ装置におけるパリティ格納方法において、

更新されたパリティ群をシーケンシャルにまとめて書き込む前記パリティ用のディスクを複数の領域に分割し、更新される前のパリティが所属する領域ごとに、更新されたパリティ群を作成し、該作成されたパリティ群を該パリティ群が所属する領域に、かつ書き込み要求の発行順に更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

## 3

【請求項10】 上位装置に接続され、キャッシュメモリと少なくとも1台の制御装置を備えるディスクアレイコントローラと、該ディスクアレイコントローラに接続され複数台のディスク装置からなる少なくとも1つの論理グループを備えるディスクアレイユニットとを備え、前記ディスクアレイコントローラの制御装置が、前記上位装置から1回に読み出したまたは書き込みする単位で転送されてきたデータを分割せずに前記複数台のディスク装置のいずれかに格納し、前記複数台のデータ用のディスク装置に格納されている各データによりパリティを生成し、この生成したパリティを、前記複数台のディスク装置のいずれかに格納するディスクアレイ装置におけるパリティ格納方法であって、

前記複数台のディスク装置の夫々にパリティ格納領域を設け、該パリティ格納領域が割り当てられるディスク装置内の領域を各ディスク装置において夫々異なるディスク装置内の領域とし、

このディスクアレイ装置に対し上位装置から発行された書き込み要求に対し、前記制御装置はパリティを更新した後、この更新したパリティを前記キャッシュメモリに格納し、同様に上位装置から発行されてきた別の書き込み要求に対し更新したパリティも該キャッシュメモリに格納し、これらの更新されたパリティを該キャッシュメモリ内に貯蔵し、このキャッシュメモリ内の更新されたパリティ群を更新される前のパリティが所属するディスク装置のパリティ格納領域毎に作成し、対応するディスク装置のパリティ格納領域に、該パリティ群を書き込み要求の発行順に更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項11】 請求項1乃至請求項3のいずれかの請求項記載のディスクアレイ装置におけるパリティ格納方法において、

前記パリティ用のディスク装置をフラッシュメモリとし、該パリティ用のフラッシュメモリ上の前記更新前のパリティ群を消去した後、前記キャッシュメモリ内の更新されたパリティ群を書き込み要求の発行順に前記パリティ用のフラッシュメモリの消去された更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項12】 請求項11記載のディスクアレイ装置におけるパリティ格納方法において、

前記キャッシュメモリにアドレス変換用テーブルを設け、

該テーブルには、パリティグループの各データの論理アドレスと該論理アドレスに対応するディスク装置番号およびディスク装置内アドレスと、パリティの論理アドレスと該論理アドレスに対応するフラッシュメモリチップ番号およびフラッシュメモリチップ内アドレスとパ

## 4

リティをキャッシュメモリ内に貯蔵した場合のキャッシュアドレスとキャッシュメモリ内に貯蔵されたパリティとディスク装置内に格納されているパリティの何れが有効を示すフラグを備えることを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項13】 請求項12記載のディスクアレイ装置におけるパリティ格納方法において、

更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む時の、実際にパリティをフラッシュメモリに書き込む前に、フラッシュメモリ内のパリティを書き込むアドレスの消去を行う際に、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれているフラッシュメモリのアドレスにおいては消去を行わずに、書き込み要求が発行され、すでに更新されているフラッシュメモリの無効なパリティのアドレスに対してのみ、消去を行うことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項14】 請求項13記載のディスクアレイ装置におけるパリティ格納方法において、

更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む際に、書き込み要求が発行され、すでに消去されているフラッシュメモリの無効なパリティのアドレスに対してのみ、書き込みを行うことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項15】 請求項12記載のディスクアレイ装置におけるパリティ格納方法において、

更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む際、フラッシュメモリの更新されたパリティ群の書き込み先の領域に、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれている場合、この有効なパリティをキャッシュメモリに読み出し、更新されたパリティと合わせてパリティ群を構成すると共にパリティの書き込み順を決定し、該パリティ群に対応するフラッシュメモリ上の書き込み前のパリティ群を消去した後、前記キャッシュメモリ内のパリティ群を前記書き込み順にフラッシュメモリの書き込み前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項16】 請求項12または請求項15記載のディスクアレイ装置におけるパリティ格納方法において、更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む際、フラッシュメモリ内の低いアドレスから高いアドレスに向かって順次書き込んでいき、最も高いアドレスまでパリティを書き込んだら、最も低いアドレスに戻り、又、順次パリティを書き込んでいく、リングバッファのようにフラッシュメモリにパリティをシーケンシャルに書き込んでいくことを特徴とするディスクアレイ装置におけるパリティ格納方法。

## 5

【請求項17】 請求項16記載のディスクアレイ装置におけるパリティ格納方法において、

フラッシュメモリ内の最も低いアドレスにパリティを書き込んだ回数のカウントをすることを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項18】 請求項17記載のディスクアレイ装置におけるパリティ格納方法において、

前記カウントした回数に基づきフラッシュメモリの寿命を判定し、寿命がきたことを出力することを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項19】 請求項11記載のディスクアレイ装置におけるパリティ格納方法において、

前記パリティ用のフラッシュメモリを複数のフラッシュメモリチップで構成し、更新されたパリティ群をまとめてフラッシュメモリに書き込む際、複数のフラッシュメモリチップに、更新されたパリティ単位に、更新前のパリティを消去した後、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項20】 請求項11記載のディスクアレイ装置におけるパリティ格納方法において、

前記パリティ用のフラッシュメモリを複数のフラッシュメモリチップで構成し、更新されたパリティ群をまとめてフラッシュメモリに書き込む際、複数のフラッシュメモリチップに、バイト単位に、更新前のパリティを消去した後、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込むことを特徴とするディスクアレイ装置におけるパリティ格納方法。

【請求項21】 上位装置に接続され、キャッシュメモリと少なくとも1台の制御装置を備えるディスクアレイコントローラと、該ディスクアレイコントローラに接続され複数台のディスク装置と複数個のフラッシュメモリからなる少なくとも1つの論理グループを備えるディスクアレイユニットとを備え、

前記ディスクアレイコントローラの制御装置が、前記上位装置から1回に読み出しまたは書込みする単位で転送されてきたデータを分割せずに前記複数台のディスク装置の内の複数台のデータ用のディスク装置のいずれかに格納し、前記複数台のデータ用のディスク装置に格納されている各データによりパリティを生成し、この生成したパリティを、前記フラッシュメモリに格納するディスクアレイ装置において、

前記制御装置は、このディスクアレイ装置に対し上位装置から発行された書き込み要求に対し、パリティを更新した後、この更新したパリティを前記キャッシュメモリに格納し、同様に上位装置から発行されてきた別の書き込み要求に対し更新したパリティも該キャッシュメモリに格納し、これらの更新されたパリティを該キャッシュメモリ内に貯蔵する手段と、このキャッシュメモリ内の

## 6

更新されたパリティ群を、該パリティ群に対応するフラッシュメモリ上の更新前のパリティ群を消去した後、前記キャッシュメモリ内の更新されたパリティ群を書き込み要求の発行順に前記パリティ用のフラッシュメモリの消去された更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込まれたディスクアレイ装置において、データが格納されているドライブに障害が発生した場合は、このドライブに格納されているデータを回復するたびに、パリティが格納されているフラッシュメモリから、当該パリティを読み出すことを特徴とするディスクアレイにおける障害回復方法。

【請求項22】 上位装置に接続され、キャッシュメモリと少なくとも1台の制御装置を備えるディスクアレイコントローラと、該ディスクアレイコントローラに接続され複数台のディスク装置からなる少なくとも1つの論理グループを備えるディスクアレイユニットとを備え、前記ディスクアレイコントローラの制御装置が、前記上位装置から1回に読み出しまたは書込みする単位で転送されてきたデータを分割せずに前記複数台のディスク装置の内の複数台のデータ用のディスク装置のいずれかに格納し、前記複数台のデータ用のディスク装置に格納されている各データによりパリティを生成し、この生成したパリティを、前記複数台のディスク装置の内のパリティ用のディスク装置に格納するディスクアレイ装置において、

前記制御装置は、このディスクアレイ装置に対し上位装置から発行された書き込み要求に対し、パリティを更新した後、この更新したパリティを前記キャッシュメモリに格納し、同様に上位装置から発行されてきた別の書き込み要求に対し更新したパリティも該キャッシュメモリに格納し、これらの更新されたパリティを該キャッシュメモリ内に貯蔵する手段と、このキャッシュメモリ内の更新されたパリティ群を書き込み要求の発行順に前記パリティ用のディスク装置の更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項23】 請求項22記載のディスクアレイ装置において、

前記ディスクアレイコントローラに更新されたパリティを格納する専用メモリを設け、更新されたパリティを前記キャッシュメモリに替えて前記専用メモリに格納するようにしたことを特徴とするディスクアレイ装置。

【請求項24】 請求項23記載のディスクアレイ装置において、

前記専用メモリを揮発メモリとしたことを特徴とするディスクアレイ装置。

【請求項25】 請求項22記載のディスクアレイ装置において、

前記キャッシュメモリにアドレス変換用テーブルを設け、

該テーブルには、パリティグループの各データの論理アドレスと該論理アドレスに対応するディスク装置番号およびディスク装置内アドレスと、パリティの論理アドレスと該論理アドレスに対応するディスク装置番号およびディスク装置内アドレスとパリティをキャッシュメモリ内に貯蔵した場合のキャッシュアドレスとキャッシュメモリ内に貯蔵されたパリティとディスク装置内に格納されているパリティの何れが有効かを示すフラグを備えることを特徴とするディスクアレイ装置。

【請求項 26】 請求項 25 記載のディスクアレイ装置において、

前記制御装置は、更新されたパリティ群をシーケンシャルにまとめてディスク装置に書き込む途中において、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれているディスク装置のアドレスにおいては、一旦書き込みを中断し、書き込み要求が発行され、すでに更新されているディスク装置の無効なパリティのアドレスにおいて、シーケンシャル書き込みを再開する手段を備えることを特徴とするディスクアレイ装置。

【請求項 27】 請求項 25 記載のディスクアレイ装置におけるパリティ格納方法において、

前記制御装置は、更新されたパリティ群をシーケンシャルにまとめてディスク装置に書き込む際、ディスク装置の書き込み先の領域に、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれている場合、この有効なパリティをキャッシュメモリに読み出し、更新されたパリティと合わせてパリティ群を構成すると共にパリティの書き込み順を決定する手段と、該パリティ群を該書き込み順にディスク装置の書き込み前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 28】 請求項 25 記載のディスクアレイ装置において、

前記アドレス変換用テーブルには、不必要になったために削除されたデータであることを示す無効フラグを前記各データ対応に備えることを特徴とするディスクアレイ装置。

【請求項 29】 請求項 28 記載のディスクアレイ装置において、

前記制御装置は、ディスク装置に新規にデータを書き込む際、ディスク装置に未使用な領域がない場合に、前記無効フラグが立っているデータの領域にデータを書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 30】 請求項 22 記載のディスクアレイ装置において、

前記パリティ用のディスク装置を複数台設け、前記制御装置は、更新されたパリティ群をまとめてディスク装置に書き込む際、複数のパリティ用のディスク装置に、更

新されたパリティ単位に、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 31】 請求項 22 記載のディスクアレイ装置において、

前記パリティ用のディスク装置を複数台設け、前記制御装置は、更新されたパリティ群をまとめてディスク装置に書き込む際、複数のパリティ用のディスク装置に、バイト単位に、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 32】 請求項 22 記載のディスクアレイ装置において、

更新されたパリティ群をシーケンシャルにまとめて書き込む前記パリティ用のディスクを複数の領域に分割し、前記制御装置は、更新される前のパリティが所属する領域ごとに、更新されたパリティ群を作成する手段と、該作成されたパリティ群を該パリティ群が所属する領域に、かつ書き込み要求の発行順に更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 33】 上位装置に接続され、キャッシュメモリと少なくとも 1 台の制御装置を備えるディスクアレイコントローラと、該ディスクアレイコントローラに接続され複数台のディスク装置からなる少なくとも 1 つの論理グループを備えるディスクアレイユニットとを備え、前記ディスクアレイコントローラの制御装置が、前記上位装置から 1 回に読み出しまたは書き込みする単位で転送されてきたデータを分割せずに前記複数台のディスク装置のいずれかに格納し、前記複数台のデータ用のディスク装置に格納されている各データによりパリティを生成し、この生成したパリティを、前記複数台のディスク装置のいずれかに格納するディスクアレイ装置において、前記複数台のディスク装置の夫々にパリティ格納領域を設け、該パリティ格納領域が割り当てられるディスク装置内の領域を各ディスク装置において夫々異なるディスク装置内の領域とし、

前記制御装置は、このディスクアレイ装置に対し上位装置から発行された書き込み要求に対し、パリティを更新した後、この更新したパリティを前記キャッシュメモリに格納し、同様に上位装置から発行されてきた別の書き込み要求に対し更新したパリティも該キャッシュメモリに格納し、これらの更新されたパリティを該キャッシュメモリ内に貯蔵し、このキャッシュメモリ内の更新されたパリティ群を更新される前のパリティが所属するディスク装置のパリティ格納領域毎に作成する手段と、該パリティ群を対応するディスク装置のパリティ格納領域に、かつ書き込み要求の発行順に更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 3 4】 請求項 2 2 乃至請求項 2 4 のいずれかの請求項記載のディスクアレイ装置において、前記パリティ用のディスク装置をフラッシュメモリとし、該パリティ用のフラッシュメモリ上の前記更新前のパリティ群を消去した後、前記キャッシュメモリ内の更新されたパリティ群を書き込み要求の発行順に前記パリティ用のフラッシュメモリの消去された更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 3 5】 請求項 3 4 記載のディスクアレイ装置において、

前記キャッシュメモリにアドレス変換用テーブルを設け、

該テーブルには、パリティグループの各データの論理アドレスと該論理アドレスに対応するディスク装置番号およびディスク装置内アドレスと、パリティの論理アドレスと該論理アドレスに対応するフラッシュメモリチップ番号およびフラッシュメモリチップ内アドレスとパリティをキャッシュメモリ内に貯蔵した場合のキャッシュアドレスとキャッシュメモリ内に貯蔵されたパリティとフラッシュメモリ内に格納されているパリティの何れが有効を示すフラグを備えることを特徴とするディスクアレイ装置。

【請求項 3 6】 請求項 3 5 記載のディスクアレイ装置において、

前期制御装置は、更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む時の、実際にパリティをフラッシュメモリに書き込む前に、フラッシュメモリ内のパリティを書き込むアドレスの消去を行う際に、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれているフラッシュメモリのアドレスにおいては消去を行わずに、書き込み要求が発行され、すでに更新されているフラッシュメモリの無効なパリティのアドレスに対してのみ、消去を行う手段を備えることを特徴とするディスクアレイ装置。

【請求項 3 7】 請求項 3 6 記載のディスクアレイ装置において、

更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む際に、書き込み要求が発行され、すでに消去されているフラッシュメモリの無効なパリティのアドレスに対してのみ、書き込みを行う手段を備えることを特徴とするディスクアレイ装置。

【請求項 3 8】 請求項 3 5 記載のディスクアレイ装置において、

前記制御装置は、更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む際、フラッシュメモリの更新されたパリティ群の書き込み先の領域に、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれている場合、この有効なパリティをキャッシュメモリに読み出し、更新された

パリティと合わせてパリティ群を構成すると共にパリティの書き込み順を決定手段と、該パリティ群に対応するフラッシュメモリ上の書き込み前のパリティ群を消去する手段と、前記キャッシュメモリ内のパリティ群を前記書き込み順にフラッシュメモリの書き込み前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 3 9】 請求項 3 5 または請求項 3 8 記載のディスクアレイ装置において、

更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む際、フラッシュメモリ内の低いアドレスから高いアドレスに向かって順次書き込んでいく手段と、最も高いアドレスまでパリティを書き込んだら、最も低いアドレスに戻り、又、順次パリティを書き込んでいく、リングバッファのようにフラッシュメモリにパリティをシーケンシャルに書き込んでいく手段を備えることを特徴とするディスクアレイ装置。

【請求項 4 0】 請求項 3 9 記載のディスクアレイ装置において、

フラッシュメモリ内の最も低いアドレスにパリティを書き込んだ回数のカウントをする手段を備えることを特徴とするディスクアレイ装置。

【請求項 4 1】 請求項 4 0 記載のディスクアレイ装置において、

前記カウントした回数に基づきフラッシュメモリの寿命を判定し、寿命がきたことを出力する手段を備えることを特徴とするディスクアレイ装置。

【請求項 4 2】 請求項 3 4 記載のディスクアレイ装置において、

前記パリティ用のフラッシュメモリチップを複数台設け、前記制御装置は、更新されたパリティ群をまとめてフラッシュメモリに書き込む際、複数のフラッシュメモリチップに、更新されたパリティ単位に、更新前のパリティを消去した後、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 4 3】 請求項 3 4 記載のディスクアレイ装置において、

前記パリティ用のフラッシュメモリチップを複数台設け、前記制御装置は、更新されたパリティ群をまとめてフラッシュメモリに書き込む際、複数のフラッシュメモリチップに、バイト単位に、更新前のパリティを消去した後、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込む手段を備えることを特徴とするディスクアレイ装置。

【請求項 4 4】 上位装置に接続され、キャッシュメモリと少なくとも 1 台の制御装置を備えるディスクアレイコントローラと、該ディスクアレイコントローラに接続され複数台のディスク装置と複数個のフラッシュメモリからなる少なくとも 1 つの論理グループを備えるディス

クアレイユニットとを備え、

前記ディスクアレイコントローラの制御装置が、前記上位装置から1回に読み出したりは書き込みする単位で転送されてきたデータを分割せずに前記複数台のディスク装置の内の複数台のデータ用のディスク装置のいずれかに格納し、前記複数台のデータ用のディスク装置に格納されている各データによりパリティを生成し、この生成したパリティを、前記フラッシュメモリに格納するディスクアレイ装置において、

前記制御装置は、このディスクアレイ装置に対し上位装置から発行された書き込み要求に対し、パリティを更新した後、この更新したパリティを前記キャッシュメモリに格納し、同様に上位装置から発行されてきた別の書き込み要求に対し更新したパリティも該キャッシュメモリに格納し、これらの更新されたパリティを該キャッシュメモリ内に貯蔵する手段と、このキャッシュメモリ内の更新されたパリティ群を、一度にシーケンシャルに前記パリティ用のフラッシュメモリに書き込まれたディスクアレイ装置において、データが格納されているドライブに障害が発生した場合は、このドライブに格納されているデータを回復するたびに、パリティが格納されているフラッシュメモリから、当該パリティを読み出す手段を備えることを特徴とするディスクアレイ装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明はディスクファイルシステムに係り、特に高性能な入出力動作を可能とするディスクアレイ装置と該装置におけるパリティ格納方法および障害回復方法に関する。

【0002】

【従来の技術】現在のコンピュータシステムにおいては、CPU等の上位側が必要とするデータは2次記憶装置に格納され、CPUが必要とする時に応じ2次記憶装置に対してデータの書き込み、読み出しを行っている。この2次記憶装置としては一般に不揮発な記憶媒体が使用され、代表的なものとして磁気ディスク装置（以下ドライブとする）、光ディスクなどがあげられる。近年高度情報化に伴い、コンピュータシステムにおいて、2次記憶装置の高性能化が要求されてきた。その一つの解として、多数の比較的容量の小さなドライブにより構成されるディスクアレイが考えられている。ディスクアレイについて記載された文献として、「D. Patterson, G. Gibson, and R. H. Kartz; A Case for Redundant Arrays of Inexpensive Disks (RAID), in ACM SIGMOD Conference, Chicago, IL, (June 1988)」がある。D. Patterson, G. Gibson, and R. H. Kartz; A Case for Redundant Arrays of In

expensive Disks (RAID)において、データを分割して並列に処理を行うディスクアレイ（レベル3）とデータを分散して、独立に扱うディスクアレイ（レベル4、5）について、その性能および信頼性の検討結果が報告されている。現在この論文に書かれている方式が最も一般的なディスクアレイと考えられている。

【0003】レベル4、5のディスクアレイでは個々のデータを分割せずに独立に扱い、多数の比較的容量の小さなドライブに分散して格納するものである。以下にデータを分散して、独立に扱うディスクアレイについて説明する。レベル4はレベル5において論理グループを構成するドライブに分散しているパリティを、1台のパリティのみを格納するドライブにまとめたものである。ここで、レベル3、レベル4、レベル5について簡単に説明しておく。レベル3は、ディスクに格納するデータ#1として、例えば「001010101011・・・」を想定し、データ#1とパリティを格納するためのディスクとしてディスク#1～#5が設けられた場合、ディスク#1に「0」、ディスク#2に「0」、ディスク#3に「1」、ディスク#4に「0」を順次格納し、格納された「0010」に対するパリティをディスク#5に格納する。そして、次に同様に「1」、  
「0」、  
「1」、  
「0」を順次ディスク#1～#4に格納し、そのパリティを#5に格納してゆく。レベル4は、データとパリティを格納するためのディスクとしてディスク#1～#5が設けられた場合、データ#1、#5、・・・がディスク#1に、データ#2、#6、・・・がディスク#2に、データ#3、#7、・・・がディスク#3に、データ#4、#8、・・・がディスク#4に格納される。そして、例えば、データ#1が「01・・・」、データ#2が「00・・・」、データ#3が「11・・・」、データ#4が「00・・・」であるとすると、各データの先頭ビット「0010」に対するパリティをパリティ専用として指定されたディスク#5の先頭ビットととして格納し、以下同様に、各データの2番目のビット「1010」に対するパリティをディスク#5の2番目のビットととして格納してゆく。そして、データ#5～#8のデータ組に対するパリティデータをディスク#5に2番目のパリティデータとして格納するようにしてゆく。レベル5は、レベル4のようなパリティ専用のディスクを決めず、データ#1をディスク#1、データ#2をディスク#2、データ#3をディスク#3、データ#4をディスク#4に格納し、データ#1～#4のデータ組に対するパリティデータP1234をディスク#5に格納し、次いで、データ#5をディスク#2、データ#6をディスク#3、データ#7をディスク#4、データ#8をディスク#5に格納し、データ#5～#8のデータ組に対するパリティデータP5678をディスク#1に格納し、次いで、データ#9をディスク#1、

データ#10をディスク#3、データ#11をディスク#4、データ#12をディスク#5に格納し、データ#9～#12のデータ組に対するパリティデータP9101112をディスク#2に格納するようにしてゆく。

【0004】現在、一般に使用されている汎用大型コンピュータシステムの2次記憶装置では、1ドライブ当りの容量が大きいため、他の読み出し／書き込み要求に当該ドライブが使用されて、そのドライブを使用できずに待たされることが多く発生した。上記文献に記載されたタイプのディスクアレイでは汎用大型コンピュータシステムの2次記憶装置で使用されている大容量のドライブを、多数の比較的容量の小さなドライブで構成し、データを分散して格納してあるため、読み出し／書き込み要求が増加してもディスクアレイの複数のドライブで分散して処理することが可能となり、読み出し／書き込み要求がまたされることが減少する。しかし、ディスクアレイは、このように多数のドライブにより構成されるため、部品点数が増加し障害が発生する確率が高くなる。そこで、信頼性の向上を図る目的で、パリティを用意する必要がある。このパリティによりデータを格納したドライブに障害が発生した場合、その障害ドライブ内のデータを復元することが可能となる。ディスクアレイではデータからパリティを作成しデータと同様にドライブに格納しておく。この時、パリティは、パリティの作成に関与したデータとは別のドライブに格納される。

【0005】これらのディスクアレイでは、現在一般に使用されている汎用大型コンピュータシステムと同様、2次記憶装置内では、個々のデータの格納場所（アドレス）は予め指定したアドレスに固定され、CPUから当該データへ読み出しまたは書き込みする場合、この固定されたアドレスへアクセスすることになっている。この分散して格納するディスクアレイ（レベル5）ではストレージテクノロジーコーポレーション（以下STKとする）から製品発表がされている。STK社から出願されている米国特許WO 91/20076では、レベル5の基本アーキテクチャにおいて、動的に変更可能なアドレスのテーブルを用意することにより、データ圧縮を行いデータの書き込み処理において、トラック単位で書き込み先のアドレスを動的に変換する方法について開示されている。また、IBM社の特開平4-230512号公報にはレベル5において、書き込み時に書き込むデータと、この書き込みにより更新したパリティを、それぞれ別の場所に書き込む方法について開示されている。さらに、IBM社からディスクアレイ（9337）では、レベル5においてWAD（ライト アシスト デバイス）を設けることが発表されている。

【0006】一方、近年磁気ディスクの置き換えデバイスとしてフラッシュメモリが着目されている。フラッシュメモリは不揮発な半導体メモリのため、磁気ディスクと比較して高速にデータの読み出し、書き込みが可能で

ある。しかし、フラッシュメモリでは書き込む際に書き込み先に書き込まれているデータを消去してからでなければ書き込めない。HN28F1600シリーズのフラッシュメモリのデータシート（ADJ-203-045（A）（z））によると、データの書き込みまたは読みだし時間は、DRAM等と同様に約100ns程度だが、消去時間が10msかかる。また、フラッシュメモリでは書き込み回数に限界があり、一般にフラッシュメモリでは百万回が書き込み回数の限界とされ実用化においては問題とされている。このように、フラッシュメモリにおける、書き込み回数に限界があるという問題点を解決する方法として、書き込み時にマッピングテーブルでフラッシュメモリへの書き込回数が平均化するようにアドレス変換する方法についてIBM社から特開平5-27924号公報において開示されている。

【0007】

【発明が解決しようとする課題】現在の汎用大型計算機システム等ではドライブにより構成される2次記憶装置内では、CPUから転送されてくるデータは個々のデータの格納場所（アドレス）が予め指定したアドレスに固定され、CPUから当該データへ読み出しまたは書き込む場合は、この固定されたアドレスへアクセスすることになる。これは、ディスクアレイにおいても同じである。データを分割して並列に処理を行うディスクアレイ（レベル3）ではこのようにアドレスを固定しても影響は無いが、データを分散して、独立に扱うディスクアレイ（レベル4、5）ではアドレスを固定した場合、書き込み時に大きな処理オーバーヘッドが必要になる。この書き込み時の処理オーバーヘッドについては特願平4-230512に説明されている。以下それについて説明する。

【0008】図10は公知例で示したD. Pattersonらが提案したRAIDに述べられている、データを分散して独立に扱うディスクアレイ（レベル5）内部のデータアドレスを示している。この各アドレスにあるデータは1回の読み出し／書き込み処理される単位で、個々のデータは独立している。前述したようにこのようなシステムでは、信頼性を向上するためパリティを設定することが不可欠である。本システムでは各ドライブ内の同一アドレスのデータによりパリティが作成される。すなわち、ドライブ#1から4までのアドレス（1, 1）のデータの組によりパリティが作成され、パリティを格納するドライブの（1, 1）に格納される。本システムでは読み出し／書き込み処理は現在の汎用大型計算機システムと同様に各ドライブに対し当該データをアクセスする。このようなディスクアレイにおいて、例えばドライブ#3のアドレス（2, 2）のデータを更新する場合、まず、更新される前のドライブ#3の（2, 2）のデータと、パリティを格納してあるドライブの（2, 2）のパリティを読み出す（1）。読み出したデータと、読み出したパリティと、更新する新しいデータとの排他的論

理和をとり、新たなパリティを作成する(2)。パリティの作成完了後、更新する新しいデータをドライブ#3の(2, 2)に、新パリティをパリティを格納するドライブの(2, 2)に格納する(3)。

【0009】このようなレベル5のディスクアレイでは、データの格納されているドライブ、パリティの格納されているドライブから古いデータとパリティを読みだすため、ディスクを平均1/2回転待ち、それから読み出してパリティを作成する。この新しく作成したパリティを書き込むため更に1回転必要となり、データを書き替える場合最低で1.5回転待たなければならない。特に更新されたパリティを書き込む際に待たされる1回転の回転待ちが、書き込み時の性能低下を引き起こす大きな問題である。このように、ドライブにおいては1.5回転ディスクの回転を待つということは非常に大きなオーバーヘッドとなる。この書き込み時のオーバーヘッドはレベル4においても同様である。このような書き込み時のオーバーヘッドを削減するため、書き込み先のアドレスを動的に変換する方法が考えられ、STK社から出願されているWO 91/20076に開示されている。また、IBM社から出願されている特願平4-230512号公報においても、書き込み時において書き込みデータを書き込みデータが書き込まれるアドレスではなく別のアドレスに書き込むことにより書き込みオーバーヘッドを削減する方法について開示されている。このように、レベル5のディスクアレイでは、読み出しと比較し書き込み時ではパリティ生成とこの生成したパリティを書き込む処理のオーバーヘッドが非常に大きいため、CPUからの読み出し、書き込み要求が多いときには、この処理オーバーヘッドが性能低下の大きな原因となる。

【0010】本発明の目的は、ディスクアレイにおけるパリティを書き込み処理のオーバーヘッドを大幅に減少させることにある。本発明の他の目的は、ディスクアレイにおいて、パリティをフラッシュメモリ(FMEM)に格納することで、書き込み処理のオーバーヘッドを大幅に減少させることにある。本発明のさらに他の目的としては、パリティを格納するFMEMの書き込み回数を平均化させることにある。

【0011】

【課題を解決するための手段】上位装置から1回に読み出しまたは書き込みする単位で転送されてきたデータを分割せずに複数台のディスク装置の内の複数台のデータ用のディスク装置のいずれかに格納し、複数台のデータ用のディスク装置に格納されている各データによりパリティを生成し、この生成したパリティを、複数台のディスク装置の内のパリティ用のディスク装置に格納するディスクアレイ装置において、このディスクアレイ装置に対し上位装置から発行された書き込み要求に対し、パリティを更新した後、この更新したパリティを前記キャッシュメモリに格納し、同様に上位装置から発行されてきた

別の書き込み要求に対し更新したパリティも該キャッシュメモリに格納し、これらの更新されたパリティを該キャッシュメモリ内に貯蔵し、このキャッシュメモリ内の更新されたパリティ群を書き込み要求の発行順に前記パリティ用のディスク装置の更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むようにしている。また、前記キャッシュメモリにアドレス変換用テーブルを設け、該テーブルには、パリティグループの各データの論理アドレスと該論理アドレスに対応するディスク装置番号およびディスク装置内アドレスと、パリティの論理アドレスと該論理アドレスに対応するディスク装置番号およびディスク装置内アドレスとパリティをキャッシュメモリ内に貯蔵した場合のキャッシュアドレスとキャッシュメモリ内に貯蔵されたパリティとディスク装置内に格納されているパリティの何れが有効かを示すフラグを備えるようにしている。また、キャッシュメモリにアドレス変換用テーブルを設けておき、更新されたパリティ群をシーケンシャルにまとめてディスク装置に書き込む途中において、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれているディスク装置のアドレスにおいては、一旦書き込みを中断し、書き込み要求が発行され、すでに更新されているディスク装置の無効なパリティのアドレスにおいて、シーケンシャル書き込みを再開するようにしている。また、キャッシュメモリにアドレス変換用テーブルを設けておき、更新されたパリティ群をシーケンシャルにまとめてディスク装置に書き込む際、ディスク装置の書き込み先の領域に、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれている場合、この有効なパリティをキャッシュメモリに読み出し、更新されたパリティと合わせてパリティ群を構成すると共にパリティの書き込み順を決定し、該パリティ群を該書き込み順にディスク装置の書き込み前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むようにしている。また、パリティ用のディスク装置を複数台設け、更新されたパリティ群をまとめてディスク装置に書き込む際、複数のパリティ用のディスク装置に、更新されたパリティ単位に、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込むようにしている。更新されたパリティ群をシーケンシャルにまとめて書き込むパリティ用のディスクを複数の領域に分割し、更新される前のパリティが所属する領域ごとに、更新されたパリティ群を作成し、該作成されたパリティ群を該パリティ群が所属する領域に、かつ書き込み要求の発行順に更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むようにしている。また、パリティ用のディスク装置を設けず、複数台のディスク装置の夫々にパリティ格納領域を設け、該パリティ格納領域が割り当てられるディスク装置内の領域を各ディスク装置において夫々異なるディスク装置内の

領域とし、上位装置から発行された書き込み要求に対し、パリティを更新した後、この更新したパリティをキャッシュメモリに格納し、同様に上位装置から発行されてきた別の書き込み要求に対し更新したパリティも該キャッシュメモリに格納し、これらの更新されたパリティを該キャッシュメモリ内に貯蔵し、このキャッシュメモリ内の更新されたパリティ群を更新される前のパリティが所属するディスク装置のパリティ格納領域毎に作成し、対応するディスク装置のパリティ格納領域に、該パリティ群を書き込み要求の発行順に更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むようにしている。

【0012】また、上位装置から1回に読み出しまたは書き込みする単位で転送されてきたデータを分割せずに複数台のディスク装置の内の複数台のデータ用のディスク装置のいずれかに格納し、複数台のデータ用のディスク装置に格納されている各データによりパリティを生成し、この生成したパリティを、複数台のディスク装置の内のパリティ用のディスク装置に格納するディスクアレイ装置において、該パリティ用のディスク装置をフラッシュメモリとし、該パリティ用のフラッシュメモリ上の前記更新前のパリティ群を消去した後、前記キャッシュメモリ内の更新されたパリティ群を書き込み要求の発行順に前記パリティ用のフラッシュメモリの消去された更新前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むようにしている。また、キャッシュメモリにアドレス変換用テーブルを設け、該テーブルには、パリティグループの各データの論理アドレスと該論理アドレスに対応するディスク装置番号およびディスク装置内アドレスと、パリティの論理アドレスと該論理アドレスに対応するフラッシュメモリチップ番号およびフラッシュメモリチップ内アドレスとパリティをキャッシュメモリ内に貯蔵した場合のキャッシュアドレスとキャッシュメモリ内に貯蔵されたパリティとディスク装置内に格納されているパリティの何れが有効を示すフラグを備えるようにしている。また、更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む時の、実際にパリティをフラッシュメモリに書き込む前に、フラッシュメモリ内のパリティを書き込むアドレスの消去を行う際に、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれているフラッシュメモリのアドレスにおいては消去を行わずに、書き込み要求が発行され、すでに更新されているフラッシュメモリの無効なパリティのアドレスに対してのみ、消去を行うようにしている。さらに、更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む際に、書き込み要求が発行され、すでに消去されているフラッシュメモリの無効なパリティのアドレスに対してのみ、書き込みを行うようにしている。また、更新されたパリティ群をシーケンシャルにまとめて

フラッシュメモリに書き込む際、フラッシュメモリの更新されたパリティ群の書き込み先の領域に、書き込み要求が発行されていないため更新されていない有効なパリティが書き込まれている場合、この有効なパリティをキャッシュメモリに読み出し、更新されたパリティと合わせてパリティ群を構成すると共にパリティの書き込み順を決定し、該パリティ群に対応するフラッシュメモリ上の書き込み前のパリティ群を消去した後、前記キャッシュメモリ内のパリティ群を前記書き込み順にフラッシュメモリの書き込み前のパリティ群の一連のアドレスへ一度にシーケンシャルに書き込むようにしている。また、更新されたパリティ群をシーケンシャルにまとめてフラッシュメモリに書き込む際、フラッシュメモリ内の低いアドレスから高いアドレスに向かって順次書き込んでいき、最も高いアドレスまでパリティを書き込んだら、最も低いアドレスに戻り、又、順次パリティを書き込んでいく、リングバッファのようにフラッシュメモリにパリティをシーケンシャルに書き込んでいくようにしている。また、フラッシュメモリ内の最も低いアドレスにパリティを書き込んだ回数のカウントをするようにしている。また、上記カウントした回数に基づきフラッシュメモリの寿命を判定し、寿命がきたことを出力するようにしている。また、パリティ用のフラッシュメモリを複数のフラッシュメモリチップで構成し、更新されたパリティ群をまとめてフラッシュメモリに書き込む際、複数のフラッシュメモリチップに、更新されたパリティ単位に、更新前のパリティを消去した後、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込むようにしている。また、パリティ用のフラッシュメモリを複数のフラッシュメモリチップで構成し、更新されたパリティ群をまとめてフラッシュメモリに書き込む際、複数のフラッシュメモリチップに、バイト単位に、更新前のパリティを消去した後、書き込み要求の発行順に更新前のパリティの一連のアドレスへ一度に並列に書き込むようにしている。また、データが格納されているドライブに障害が発生した場合は、このドライブに格納されているデータを回復するたびに、パリティが格納されているフラッシュメモリから、当該パリティを読み出し、障害回復をするようにしている。

#### 【0013】

【作用】データの書き込みによるパリティの更新後、更新された新パリティを更新のたびに更新前のパリティが格納されていたドライブのアドレスに書き込むと、その都度回転待ちが必要となる。本発明では、更新パリティをキャッシュメモリに貯蔵しておき、これらの更新パリティをシーケンシャルにまとめ書きすることにより、このまとめ書きを行なう時に0.5回転の回転待ちが必要となるが、まとめ書きを開始以降は回転待ちは無い。つまり、まとめられた更新パリティの集団の中で、一番最初に書き込まれるパリティについては0.5回転の回転

待ちが必要であるが、引き続き書き込まれる2番目以降のパリティについては回転待ちがなくなる。また、更新パリティをキャッシュメモリに貯蔵しておき、これらの更新パリティをシーケンシャルにFMEMにまとめ書きすることにより、このまとめ書きを行なう時は最初に一括消去の時間(約10ms)が必要となるのみである。FMEMでは部分消去時間と一括消去時間はほとんど同じである。つまり、まとめられた更新パリティの集団の中で、一番最初に書き込まれるパリティについては、全ての更新パリティに対応する更新前のパリティを一括消去する時間が必要であるが、引き続き書き込まれる2番目以降のパリティについてはFMEMへの書き込み時間のみとなる。また、FMEMに対する新パリティのシーケンシャルなまとめ書きは、アドレスの低い方から高い方へ方向に行なわれるため、FMEMチップでは書き込み回数が平均化される。

【0014】

【実施例】

(実施例1) 以下本発明の一実施例を説明する。図1は、本実施例のハードウェア構成を示し、1はCPU、2はディスクアレイコントローラ(以下ADC)、3はディスクアレイユニット(以下ADU)である。ADU3は複数の論理グループ10により構成され、個々の論理グループ10はm台のドライブ12と、各々のドライブ12とADC2を接続するディスクアレイユニットバス9-1から9-4により構成される。なお、このドライブ12の数については特に制限は無い。この論理グループ10は障害回復単位で、この論理グループ10内の各ドライブ12内の各データによりパリティを作成する。本実施例ではm-1台の個々のドライブ12内のデータから各々のパリティが作成される。

【0015】次にADC2の内部構造について図1を用いて説明する。ADC2はチャンネルバスディレクタ5と2個のクラスタ13とバッテリバックアップ等により不揮発化された半導体メモリであるキャッシュメモリ7により構成される。このキャッシュメモリ7にはデータとアドレス変換用テーブルが格納されている。このキャッシュメモリ7およびその中のアドレス変換用テーブルはADC2内の全てのクラスタ13において共有で使用される。クラスタ13はADC2内において独立に動作可能なバスの集合で、各クラスタ13間においては電源、回路は全く独立となっている。クラスタ13はチャンネル、キャッシュメモリ7間のバスであるチャンネルバス6と、キャッシュメモリ7、ドライブ12間のバスであるドライブバス6-1から4が、それぞれ、2個ずつで構成されている。それぞれのチャンネルバス6-1から4とドライブバス8はキャッシュメモリ7を介して接続されている。CPU1より発行されたコマンドは外部インターフェースバス4を通してADC2のチャンネルバスディレクタ5に発行される。ADC2は2個のクラスタ13

により構成され、それぞれのクラスタは2個のバスで構成されるため、ADC2は合計4個のバスにより構成される。このことから、ADC2ではCPU1からのコマンドを同時に4個まで受け付けることが可能である。そこで、CPU1からコマンドが発行された場合ADC2内のチャンネルバスディレクタ5によりコマンドの受付が可能かどうか判断する。

【0016】図2は図1のチャンネルバスディレクタ5と1クラスタ13-1内の内部構造を示した図である。図2に示すように、CPU1からADC2に送られてきたコマンドはインターフェースアダプタ(以下IF Adp)15により取り込まれ、マイクロプロセッサであるMP20はクラスタ内の外部インターフェースバス4の中で使用可能なバスがあるかを調べ、使用可能な外部インターフェースバス4がある場合はMP20はチャンネルバススイッチ16を切り換えてコマンドの受け付け処理を行ない、受け付けられない場合は受付不可の応答をCPU1へ送る。

【0017】(アドレス変換法) 本実施例ではADU3を構成するドライブ12はSCSIインターフェースのドライブを使用する。以下にアドレス変換について説明する。CPU1は論理アドレスとしてデータ名を指定し、ADC2のMP20により実際のドライブ12内の物理的地址であるドライブ12内のアドレス(SCSI内Addr)に変換される。SCSI内Addrは図11に示すように当該データが格納されているトラックが所属するシリンダの位置とそのシリンダ内において当該データが格納されているトラックを決定するヘッドアドレスと、そのトラック内のレコードの位置で構成される。具体的には要求データが格納されている当該ドライブ12の番号と当該ドライブ12内のシリンダ番号であるシリンダアドレスとシリンダ内においてトラックを選択するヘッドの番号であるヘッドアドレスとレコードアドレスからなる。このアドレス変換には以下に示すようなアドレス変換用のテーブル(以下アドレステーブルとする)が使用される。ADC2内のキャッシュメモリ7には、その内部の適当な領域に図4に示すようなアドレステーブルが格納されている。

【0018】アドレステーブルは論理アドレス27に対し、無効データの場合オン(1)となる無効フラグ28と、データが格納されているドライブ12のアドレスであるデータドライブ番号29(DDrive No.)と、そのドライブ12内の実際にデータが格納されている物理アドレスであるSCSI内Addr30と、そのデータがキャッシュメモリ7内にある場合のキャッシュメモリ7内のアドレスであるキャッシュアドレス31と、キャッシュメモリ7内にデータがある場合オン

(1)とするキャッシュフラグ32と、パリティグループにおいてパリティの論理アドレスであるP論理アドレス33と、パリティが格納されているドライブ番号であ

るパリティドライブ番号34 (P Drive No.) と、そのドライブ12内の実際にパリティが格納されている物理アドレスであるP SCSI内Addr35と、パリティの格納されているキャッシュアドレスであるP キャッシュアドレス36と、パリティがキャッシュメモリ7内に存在するか否かを示すPキャッシュフラグ37により構成されている。パリティは、論理グループ10を構成する各ドライブ12において同一SCSI内Addr30のデータにより作成される。パリティグループは、論理グループ10を構成する各ドライブ12において同一SCSI内Addr30のデータと、これらのデータから作成されたパリティにより構成される。具体的には、図4においてSCSI内Addr30がDADR1についてはSD#1のドライブ12に格納されているData#1と、SD#2のドライブ12に格納されているData#2と、SD#3のドライブ12に格納されているData#3と、SD#4のドライブ12に格納されているData#4とにより作成されたパリティであるParity#1がSD#5のドライブ12のP SCSI内ADDRがDADR5に格納され、これらのデータとパリティがパリティグループを構成する。

【0019】以上に説明したアドレステーブルにより、CPUが指定する論理アドレス27に基づき、データが実際に格納されているドライブ番号29とSCSI内Addr30を決定する。例えば、図4においてCPU1からData#2に対し要求を発行してきた場合、アドレステーブルからSD#2のドライブ12内のSCSI内Addr30としてDADR1が該当していることが分かり、物理的なアドレスへ変換される。また、このData#2に対応するパリティは、P論理アドレス33がParity#1で、パリティドライブ番号(P Drive No) 34がSD#5のドライブ12であり、P SCSI内ADDRがDADR5の位置に格納されている。このように、CPU1から指定された論理アドレス27を、実際に読み出し/書き込みを行うドライブ12の物理的なアドレスに変換した後、SD#2のドライブ12のData#2に対し読み出しまたは書き込み要求が発行される。この時アドレステーブルにおいてData#2の論理アドレス27ではキャッシュフラグ32がオン(1)のため、このデータはキャッシュメモリ7内のCADR5に存在する。もし、キャッシュフラグ32がオフ(0)であればキャッシュメモリ7内には、当該データは存在しない。この、アドレステーブルはシステムの電源をオンした時に、MP20により論理グループ10内のある特定のドライブ12から、キャッシュメモリ7にCPU1の関知無しに自動的に読み込まれる。一方、電源をオフする時はMP20によりキャッシュメモリ7内のアドレステーブルを、読み込んできたドライブ12内の所定の場所にCPU1の関知無しに自動的に格納する。

【0020】(読み出し処理)次に、ADC2内での具体的なI/O処理について図1、図2を用いて説明する。CPU1より発行されたコマンドはIF Adp15を介してADC2に取り込まれ、MP20により読み出し要求か書き込み要求か解読される。まず、読み出し要求の場合の処理方法を以下に示す。MP20が読み出し要求のコマンドを認識すると、MP20はCPU1から送られてきた論理アドレスをアドレステーブルを参照し、キャッシュメモリ7内に存在するかどうかキャッシュフラグ32を調べ、判定する。キャッシュフラグ32がオンでキャッシュメモリ7内に格納されている場合(キャッシュヒット)は、MP20がキャッシュメモリ7から当該データを読み出す制御を開始し、キャッシュメモリ7内に無い場合(キャッシュミス)は当該ドライブ12へその内部の当該データを読みだす制御を開始する。キャッシュヒット時はMP20はアドレステーブルによりCPU1から指定してきた論理アドレス27に対し、当該データが格納されているキャッシュメモリ7のキャッシュアドレス31に変換し、キャッシュメモリ7へ当該データを読み出しに行く。具体的にはMP20の指示の元でキャッシュアダプタ回路(C Adp)23によりキャッシュメモリ7から当該データは読み出される。C Adp23はキャッシュメモリ7に対するデータの読み出し、書き込みをMP20の指示で行う回路で、キャッシュメモリ7の状態の監視、各読み出し、書き込み要求に対し排他制御を行う回路である。C Adp23により読み出されたデータはデータ制御回路(DCC)22の制御によりチャンネルインターフェース回路(CH IF)21に転送される。CH IF21ではCPU1におけるチャンネルインターフェースの protocols に変換し、チャンネルインターフェースに対応する速度に速度調整する。具体的にはCPU1、ADC2間のチャンネルインターフェースを光のインターフェースにした場合、光のインターフェースの protocols をADC2内では電気処理での protocols に変換する。CH IF21における protocols 変換および速度調整後は、チャンネルバスディレクタ5において、チャンネルバススイッチ16が外部インターフェースバス4を選択しIF Adp15によりCPU1へデータ転送を行なう。

【0021】一方、キャッシュミス時はキャッシュヒット時と同様にアドレステーブルにより、CPU1が指定した論理アドレス27から当該ドライブ番号とそのドライブ12内の実際にデータが格納されているSCSI内Addr30を認識し、そのアドレスに対し、MP20はDrive IF24に対し、当該ドライブ12への読み出し要求を発行するように指示する。Drive IF24ではSCSIの読み出し処理手順に従って、読み出しコマンドをドライブユニットバス9-1または9-2を介して発行する。Drive IF24から読み出しコマンドを発行された当該ドライブ12においては

指示された SCS I 内 Addr 30 ヘシーク、回転待ちのアクセス処理を行なう。当該ドライブ 12 におけるアクセス処理が完了した後、当該ドライブ 12 は当該データを読み出しドライブユニットバス 9 を介して Drive IF 24 へ転送する。Drive IF 24 では転送されてきた当該データをドライブ 12 側のキャッシュアダプタ回路 (C Adp) 14 に転送し、(C Adp) 14 ではキャッシュメモリ 7 にデータを格納する。この時、C Adp 14 はキャッシュメモリ 7 にデータを格納することを MP 20 に報告し、MP 20 はこの報告を元に、アドレステーブル内の CPU が読み出し要求を発行した論理アドレス 27 のキャッシュフラグ 32 をオン (1) にし、キャッシュアドレス 31 にキャッシュメモリ 7 内のデータを格納したアドレスを登録する。キャッシュメモリ 7 にデータを格納し、アドレステーブルのキャッシュフラグ 32 をオン (1) にし、キャッシュメモリ 7 内のアドレスを更新した後はキャッシュヒット時と同様な手順でキャッシュメモリ 7 からデータを読み出し、CPU 1 へ当該データを転送する。

【0022】(書き込み処理) 一方書き込み時は以下のように処理される。書き込み処理にはユーザが書き込み先の論理アドレスを指定し、そのデータを書き換える更新と、新たに空き領域に書き込む新規書き込みがある。CPU 1 から書き込み命令が発行されたとする。まず、ADC 2 の MP 20 は CPU 1 から書き込み要求のコマンドを受け取った後、コマンドを受け取った MP 20 が所属するクラス 13 内の各チャネルバス 6 において処理可能かどうかを調べ、可能な場合は処理可能だという応答を CPU 1 へ返す。CPU 1 では処理可能だという応答を受け取った後に ADC 2 へデータを転送する。この

時、ADC 2 では MP 20 の指示によりチャネルバスディレクタ 5 において、チャネルバススイッチ 16 が当該外部インターフェースバス 4 と IF Adp 15 を当該チャネルバス 6 と接続し CPU 1 と ADC 2 間の接続を確立する。CPU 1 と ADC 2 間の接続を確立後 CPU 1 からのデータ転送を受け付ける。CPU 1 から転送されてくるデータには、論理アドレスと書き込みデータ (以下新データとする) があり、これらのデータは MP 20 の指示により、CH IF 21 によりプロトコル変換を行ない、外部インターフェースバス 4 での転送速度から ADC 2 内での処理速度に速度調整する。CH IF 21 におけるプロトコル変換および速度制御の完了後、データは DCC 22 によるデータ転送制御を受け、C Adp 24 に転送され、C Adp 23 によりキャッシュメモリ 7 内に格納される。この時、CPU 1 から送られてきたデータが、論理アドレスの場合は、読み出しと同様にアドレステーブルによりアドレス変換を行い、物理アドレスに変換する。また、CPU 1 から送られてきたデータが新データの場合は、キャッシュメモリ 7 に格納したアドレスをアドレステーブル内のキャッシ

ュアドレス 31 に登録する。この時、書き込む新データをキャッシュメモリ 7 内に保持するときは、論理アドレス 27 のキャッシュフラグ 32 をオン (1) とし、保持しない場合はキャッシュフラグ 32 をオフ (0) とする。なお、キャッシュメモリ 7 内に保持されている新データに対し、さらに書き込み要求が CPU 1 から発行された場合は、キャッシュメモリ 7 内に保持されている新データを書き替える。

【0023】キャッシュメモリ 7 に格納された新データは、この新データにより新しくパリティを更新し (以下更新されたパリティを新パリティとする)、以下のように論理グループ 10 内のドライブ 12 へ新データと新パリティを格納する。まず、すでにドライブ 12 内に書き込まれているデータを新しいデータに書き換える更新の場合についてのフローを図 12 を用いて示す。本発明ではパリティは論理グループ 10 を構成するドライブ 12 において、RAID のレベル 4 のように特定のパリティ専用のドライブ 12 に格納する。

【0024】本発明の書き込み処理方法を図 3 を用いて説明する。MP 20 は CPU 1 が指定した論理アドレスからアドレステーブルを参照し、データ、パリティが格納されているドライブ 12 (DDrive No. 29, PDrive No. 34 で指定される) とそのドライブ 12 内の物理的なアドレスである SCS I 内 Addr 30, PSCS I 内 Addr 35 を認識する。図 3 に示すように CPU 1 から SD #1 のドライブ 12 の Data #1 に対し、New Data #1 に更新する書き込み要求が発行された場合、MP 20 はアドレステーブルにより更新されるデータ (旧データ) である Data #1 および更新されるパリティ (旧パリティ) である Parity #1 の物理アドレスを認識した後、それぞれのドライブに対し旧データと旧パリティの読み出しを行なう (図 3、図 12 の (1))。この時の読み出し方法は先に説明した読み出し処理におけるドライブ 12 からキャッシュメモリ 7 への読み出しと同じである。ただ、書き込み時の読み出しでは、ADC 2 の MP 20 が発行した読み出し要求のため、読み出したデータは CPU 1 へは転送せず、キャッシュメモリ 7 に転送するのみである。この様に読み出した旧データ、旧パリティと書き込む新データとで排他的論理和を行ない更新後の新パリティである New Parity #1 を作成しキャッシュメモリ 7 に格納する (図 3、図 12 の (2))。新パリティ (New Parity #1) のキャッシュメモリ 7 への格納完了後、MP 20 は新データ (New Data #1) を SD #1 のドライブ 12 の Data #1 のアドレスに書き込む (図 3、図 12 の (3))。なお、この新データの書き込みは MP 20 の管理の下で非同期に行なわれるようにしてもよい。新パリティ (New Parity #1) はキャッシュメモリ 7 にそのまま格納しておく。この時、図 4 に示すアドレステーブル

に対し論理アドレスがData #1のエントリにNew Data #1を登録し、キャッシュメモリ7に保持しておく場合はキャッシュアドレス31にキャッシュ内のアドレスを登録し、キャッシュフラグ32をオンとする。また、パリティに関してはキャッシュメモリ7に保持したままのため、Pキャッシュアドレス36にキャッシュアドレスを登録し、Pキャッシュフラグ37をオンとする。なお、この様にアドレステーブルでPキャッシュフラグ37がオンとなっているパリティは、更新済みのパリティとなり、パリティ格納用のドライブ12内に格納されているパリティは無効とされる。本発明では図5に示すように、CPU1からの新データを不揮発化されたキャッシュメモリ7内の領域に格納し、新パリティの作成が完了しキャッシュメモリ7に格納した時点で、MP20は書き込み処理を終了したとCPU1に報告する。なお、従来方法では図5に示したように、新パリティをドライブの1回転後に書き込み、MP20が書き込み処理を終了したとCPU1に報告している。新パリティのドライブ12への書き込みはMP20の管理の下で非同期に行なわれるため、ユーザからは見えない。また、新データのドライブ12への書き込みをMP20の管理の下に非同期に行なう場合は、同様にしてユーザからは見えない。以後CPU1からData #10、Data #8に対する書き込み処理が発行されれば、上記と同様に処理し、各新パリティをキャッシュメモリ7に格納していく。

【0025】キャッシュメモリ7に溜められた新パリティは、予めユーザが設定した設定値以上の新パリティがキャッシュメモリ7に溜った場合か、または、ユーザからの読み出し/書き込み要求の発行されていないタイミングが生じた場合にパリティ格納用のドライブ12にまとめて書き込む(図3、図12の(4))。このように新パリティをパリティ格納用のドライブ12にまとめて書き込む場合は、シーケンシャルに書き込まれる。この様に新パリティをパリティ格納用のドライブ12に書き込む際に、アドレステーブルのP SCSI内Addr 35に実際に新パリティを書き込んだSCSI内Addrを登録する。従来方法では制御の簡略化のため、パリティを格納するドライブ12内でパリティを格納するP SCSI内Addr 35はデータを格納するドライブ12内のデータを格納するSCSI内Addr 30と同一にしていた。しかし、本発明では新パリティはシーケンシャルにまとめて書き込まれるため、パリティ格納用のドライブ12内でパリティを格納するP SCSI内Addr 35とデータ格納用のドライブ12内のデータを格納したSCSI内Addr 30は原則として同一にはせず、異なるものになっている。

【0026】また、キャッシュメモリ7内に溜められた新パリティをシーケンシャルにまとめて書き込む最中にCPU1より読み出し要求が発行された場合、読み出し

処理にはパリティは関与しないため、先に説明したように通常の読み出し処理を行なう。一方、キャッシュメモリ7内に溜められた新パリティをシーケンシャルにまとめて書き込む最中にCPU1より書き込み要求が発行された場合は、通常の書き込み処理と同様に旧データを読み出し、旧データの読み出し後新データを書き込む。この時、キャッシュメモリ7には新データと旧データを保持し、キャッシュメモリ7内に溜められた新パリティをシーケンシャルにまとめて書き込む処理が終了次第、当該書き込まれた新パリティを旧パリティとして読み出し、この旧パリティと先の新データと旧データとにより新パリティを作成しキャッシュメモリ7に格納する。

【0027】本発明では新パリティをまとめ、シーケンシャルに書き込むが、図7、8に示すように、前に書き込まれているパリティが有効な場合、その上に新パリティを書き込んで消すわけには行かない。本発明では、アドレステーブルでPキャッシュフラグ37がオンとなっている旧パリティは、更新済みのパリティとなり、パリティ格納用のドライブ12内では無効となっているが、Pキャッシュフラグ37がオフとなっているパリティはパリティ格納用のドライブ12内においてまだ有効なパリティである。この有効なパリティが消されると、ドライブ障害が発生した場合、障害ドライブ内のこのパリティの作成に関与したデータの回復が不可能となる。以下に新パリティのシーケンシャルまとめ書きの時のパリティ格納用ドライブ12内の有効データの扱い方について説明する。図7に示すようにパリティ書き込み前のトラックにおいてパリティのP1、P2、P3はデータの書き込み要求に伴い更新された無効パリティであり、P8、P9はデータに対する書き込み要求が発行されていないため、更新されていない有効パリティである。書き込み要求1、2、3の順に書き込み要求が発行され、これによりP2、P3、P1の順に旧パリティが更新され更新済みの新パリティとしてP'2、P'3、P'1の順にキャッシュメモリ7に格納されているとする。これらの新パリティをパリティ格納用のドライブ12にシーケンシャルにまとめて書き込む場合は、旧パリティP1の位置に新パリティP'2を書き込み、有効パリティであるP8はそのまま残し、旧パリティP2の位置に新パリティP'3を書き込み、旧パリティP3の位置に新パリティP'1を順に書き込んでいく。すなわち、旧パリティの一連の位置に新パリティをその書き込み要求の発行順に順次書き込むのである。以上のように本発明ではシーケンシャルにまとめて書き込む際に、有効データはそのまま残し、飛ばして新パリティを書き込んでいく。なお、シーケンシャルにパリティを書き込んでゆく際、1つのパリティを書き込み、次のパリティの書き込みを開始するための処理をしている間に、次のパリティを書き込むためのブロックが通り過ぎてしまわないように、I/Sギャップ(これについては、“トランジスタ技術 S

PECIAL NO. 27 第20頁”に記載されている)を充分に取る必要がある。また、別の方法としては図8に示すように有効パリティであるP8、P9をMP20の指示により擬似的な読み出し要求を発行し、この擬似的な読み出し要求によりキャッシュメモリ7に読み出し、この読み出しによりMP20はアドレステーブルのPキャッシュアドレス36をセットしPキャッシュフラグ37をオンとすることで、更新する新パリティとみなし、他の新パリティと一緒にシーケンシャルにまとめて書き込む方法もある。すなわち、図8に示すように、更新する新パリティP'2、P'3、P'1と更新する新パリティとみなしたパリティP8、P9からなる更新パリティ群の書き込み順をP'2、P'3、P'1、P8、P9とし、書き込み前のパリティ群P1、P8、P2、P3、P9の一連のアドレスへ上記書き込み順にしたがって更新パリティ群を順次書き込む。書き込み結果は図8のパリティ書き込み後のトラックに示すようになる。

【0028】一方ドライブ12にすでに格納されているデータに新しいデータ追加する新規書き込みの場合は、MP20はアドレステーブルにおいて空き領域を探す。空き領域には2種類ある。まず一つはまったく使用されていない未使用領域である。この様にまったく使用されていない領域では、アドレステーブルにおいて論理アドレス27の項に論理アドレスは登録されていない。このため、MP20はアドレステーブルにおいて論理アドレス27の項に論理アドレスが登録されていない領域を探すことで、未使用領域を見つけられる。もう一つの空き領域は、以前その領域は使用されていたが(データが書き込まれていた)、ユーザがそのデータが必要でなくなったため削除した削除領域である。削除領域は、アドレステーブルにおいて論理アドレス27の項に論理アドレスが登録されてるが無効フラグ28をオン(1)としている。このため、MP20はアドレステーブルにおいて無効フラグ28がオンになっている領域を探すことで、削除領域を見つけられる。MP20が新規書き込みを行なう空き領域を決定する場合、まず、未使用領域を探す。もし、未使用領域が無い場合は削除領域を新規書き込み先に使用する。これは、未使用領域はパリティの作成に関与していない(全て0で構成されているとした)ため、新規書き込みの際のパリティの更新は、新規書き込みする新データと更新される旧パリティとの排他的論理和のみで行なえるが、削除領域のデータはユーザにとっては意味が無いデータとなっているが、パリティの作成には関与しているため、新規書き込みの際に旧データと同じように読み出して、旧パリティと新規書き込みデータとの間で排他的論理和をとり新パリティを作成しなければならない。このため、未使用領域に新規書き込みを行なうのと、削除領域に新規書き込みを行なうのでは、削除領域から削除されたデータを読み出す処理が入

らない分、未使用領域に新規書き込みを行なう方が早く処理できるためである。以上述べたようにMP20が空き領域を探し、空き領域の認識が完了した後、この空き領域に新規書き込みデータの書き込みを行ない、更新と同様にアドレステーブルに論理アドレス27を登録し、削除領域に新規書き込みを行なった場合は無効フラグ28をオフとする。以上述べたように、新規書き込みと更新では、新データの書き込み先が異なるのみで処理自体は同じである。

10 【0029】(障害回復処理)次にドライブ12に障害が発生した場合の、障害ドライブ12内のデータを回復する方法を説明する。図3に示すようにSD#1のドライブ12のData#1とSD#2のドライブ12のData#2とSD#3のドライブ12のData#3とSD#4のドライブ12のData#4からSD#5のドライブ12のParity#1が作成されている。同様にData#5、6、7、8からParity#2、Data#9、10、11、12からParity#3が作成されている。SD#1、2、3、4のドライブ12の中でどれか1台のドライブ12に障害が発生した場合、残りのドライブ12内のデータとパリティから、障害ドライブ12内のデータを回復する。本発明では、パリティはパリティを格納するドライブ12内においてランダムに格納されている。そこで、障害ドライブ12内のデータを回復する際は、MP20はSD#5のドライブ12内のパリティをキャッシュメモリ7内に全て読み出す。例えば、図3においてSD#1のドライブ12に障害が発生したとする。まず、MP20はSD#5のドライブ12からParity#1、2、3をキャッシュメモリ7に読み出し、アドレステーブルのPキャッシュアドレス36にキャッシュメモリ7内のアドレスを登録し、Pキャッシュフラグ37をオンにする。次に、MP20はSD#2、3、4のドライブ12からData#2、3、4をそれぞれ読み出し、これらのデータとアドレステーブルによりこれらのデータに対応するパリティを、アドレステーブルにより探す。当該パリティを見つけた後は、上記データと先に読み出している当該パリティであるParity#1とをパリティ生成回路(PG)25に送り、Data#1を復元する。同様にData#5、9も復元する。この様に復元したデータは、障害ドライブ12を正常なドライブ12に交換した後、この正常なドライブ12に書き込むことで回復処理を行なう。また、ドライブ12の障害時に備え予め予備の正常なドライブ12を用意してある場合は、この予備の正常なドライブ12に復元したデータを書き込みことで回復処理を行なう。

50 【0030】以上の説明では更新後の新パリティを格納するキャッシュメモリ7は不揮発な半導体メモリとした。しかし、パリティはデータとは異なり停電等によってキャッシュメモリ7から消失しても、新たに作り直す

ことが可能なため、この、新たに作成する手間を許容できるなら、キャッシュメモリ7内で旧パリティを格納する領域を揮発な半導体メモリにすることも可能である。以上の説明では、更新後の新パリティをキャッシュメモリ7に格納したが、キャッシュメモリ7ではなく専用のメモリを用意することも可能である。従来のレベル4、5では書き込み処理を行なうたびに新パリティの書き込みを行なっていたため、常にパリティの更新後に回転の回転待ちを必要としたが、本発明ではシーケンシャルなまとめ書きを行なう際の最初に0、5回転の回転待ちを必要とするのみである。

【0031】(実施例2) 本実施例では実施例1で示したように、1台のパリティ格納用のドライブ12にシーケンシャルにまとめ書きするのではなく、複数のパリティ格納用のドライブ12に対し新パリティをパラレルに書き込む方法を示す。本実施例でも実施例1と同じ処理により、データの書き込みに伴いパリティを更新し、更新した新パリティはキャッシュメモリ7に保持される。図9に示すように書き込み要求1、2、3によりData #1, #9, #8がそれぞれNew Data #1, #9, #8に更新され、このデータの更新により更新された新パリティとしてNew Parity #1, #3, #2がキャッシュメモリ7に保持されている(図9の(1)(2)(3))。実施例1と同様に予めユーザの設定値以上の新パリティがキャッシュメモリ7に溜った場合か、または、ユーザからの読み出し/書き込み要求の発行されていないタイミングで、複数のパリティ格納用のドライブ12であるSD #5, SD #6にパラレルにまとめて書き込む(図9の(5))。パラレルにまとめて書き込む単位としては、レベル3のようにバイト単位と、レベル4、5のようにパリティ単位がある。この時、各パリティ格納用のドライブ12に対する書き込み方法は、実施例1の1台のパリティ格納用のドライブ12への書き込み方法と同じである。また、本実施例の変形として、パラレルに書き込む新パリティによりパリティを作成し、SD #7のパリティ格納用のドライブ12に書き込む。この様にパリティのパリティを作成することにより、パリティ格納用のドライブ12の障害時に新たにパリティを作成する際に、データを読み出す必要が無く、その間のデータへのアクセスを受け付けることが可能となる。

【0032】(実施例3) 本実施例では、図6に示すようにパリティ格納用のドライブ12を複数の領域に分割し、各領域単位で行なう方法を説明する。この領域の分割は、SCSI内Addr 30により行なう。例えばSCSI内Addr 30がDADR1からDADRkまでを領域1とする。SD #1, 2, 3, 4のドライブ12においてSCSI内Addr 30がDADR1からDADRkまで領域1に所属する各データに対するパリティは、SD #5のドライブ12のP SCSI内Addr 3

6がDADR1からDADRkの領域1に格納される。このように、アドレステーブルにおいて、データ、パリティに対し所属する領域を対応させる。この様な領域分割を行なった場合、領域1に所属するパリティに対しデータの書き込みによるパリティの更新が行なわれた場合、新パリティは領域1のパリティとしてキャッシュメモリ7に保持する。同様にCPU1からの他の書き込みによる新パリティをキャッシュメモリ7に保持していき、領域1のパリティとして保持されている新パリティはまとめられ、領域1にシーケンシャルにまとめて書き込む。他の領域に対しても同様に各領域に所属するパリティの新パリティは、まとめてそれぞれの領域にシーケンシャルに書き込む。また、この様に領域に分割した際のドライブ12に障害が発生した場合の回復方法は、基本的には実施例1で示した領域に分割しない場合と同じである。異なるのは、障害ドライブ12内のデータを回復する際に、MP20はSD #5のドライブ12内のパリティをキャッシュメモリ7内に全て読み出さず、各領域単位に読み出す。つまり、回復処理を領域単位で行なう。

【0033】本実施例の変形例を以下に示す。本変形例では、図13に示すように、パリティを書き込む領域を、1台のパリティ書き込み用の専用ドライブ12に限定せず、論理グループ10を構成するドライブ12全体に分散させる。この様に、パリティを書き込む領域を論理グループ10を構成するドライブ12に分散させた場合と上記のように1台のドライブ12に限定した場合で異なる点を以下に示す。上記のように1台のドライブ12に限定した場合、パリティの格納先のドライブ12が限定されているため、MP20がアドレステーブルによりパリティが格納されている領域を決定する場合、SCSI内Addrのみで可能である。一方、本変形例のように、パリティを書き込む領域を論理グループ10を構成するドライブ12に分散させた場合、MP20がアドレステーブルにより領域を決定する場合、SCSI内Addrの他にドライブ番号も必要となる。この様に本変形例ではアドレス変換方法が異なるが、その他の制御方法は同じである。

【0034】(実施例4) 以下本発明の実施例4を説明する。図14は、本実施例のハードウェア構成を示し、1はCPU、2はディスクアレイコントローラ(以下ADC)、3はディスクアレイユニット(以下ADU)である。ADU3は複数の論理グループ10により構成され、個々の論理グループ10はm台のドライブ12とフラッシュメモリコントローラ(FMEMC)42と複数のフラッシュメモリチップ(FMEMチップ)40により構成されるフラッシュメモリ(FMEM)41と、各々のドライブ12またはFMEM41とADC2を接続するディスクアレイユニットバス9-1から9-4により構成される。本実施例では、各論理グループ1

0内にパリティを専用に格納するFMEM41を設け、書き込み時にADC2がレベル4の制御により作成したパリティをADC2内のキャッシュメモリ7に溜め、これらのパリティを一度にシーケンシャルにFMEM41に書き込む所に特徴がある。このように作成したパリティを溜め、まとめて一度にシーケンシャルにFMEM41に格納することで、ドライブ12で構成されたレベル4のディスクアレイで問題となった書き込み時のパリティ更新オーバーヘッド（回転待ち時間）を削減することが可能になる。なお、このドライブ12の数については特に制限は無い。この論理グループ10は障害回復単位で、この論理グループ10内の各ドライブ12内の各データによりパリティを作成する。本実施例ではADC2のMP20はm台の個々のドライブ12内のデータから各々のパリティが作成され、これらのパリティが論理グループ10の一ヶ所に集めて格納されるレベル4の制御を行う。従来のディスクアレイではこのパリティはドライブに格納されていた。

【0035】次にADC2の内部構造についての説明であるが、これは実施例1における図1および図2を用いた説明と同様であるので省略する。なお、実施例1の図2に対応する本実施例の図15ではパリティ格納用のディスクドライブがフラッシュメモリ（FMEM）41に置き換えられている。また、（アドレス変換法）については、実施例1における説明と同様であるので省略する。但し、実施例1の図4におけるPDrive No. 34、PSCSI内Addr35は図4に対応する本実施例の図17においてはFMEM34'、FMEM内Addr35'となっている。次に、（読み出し処理）についての説明は、実施例1における説明と同様であるので省略する。

【0036】（書き込み処理）次に本実施例の特徴となる書き込み時の処理について以下に示す。書き込み処理にはユーザが書き込み先の論理アドレスを指定し、そのデータを書き換える更新と、新たに空き領域に書き込む新規書き込みがある。CPU1から書き込み命令が発行されたとする。まず、ADC2のMP20はCPU1から書き込み要求のコマンドを受け取った後、コマンドを受け取ったMP20が所属するクラスタ13内の各チャンネルバス6において処理可能かどうかを調べ、可能な場合は処理可能だという応答をCPU1へ返す。CPU1では処理可能だという応答を受け取った後にADC2へデータを転送する。この時、ADC2ではMP20の指示によりチャンネルバスディレクタ5において、チャンネルバススイッチ16が当該外部インターフェースバス4とIFAdp15を当該チャンネルバス6と接続しCPU1とADC2間の接続を確立する。CPU1とADC2間の接続を確立後CPU1からのデータ転送を受け付ける。CPU1から転送されてくるデータには、論理アドレスと書き込みデータ（以下新データとする）があり、

これらのデータはMP20の指示により、CHIF21によりプロトコル変換を行ない、外部インターフェースバス4での転送速度からADC2内での処理速度に速度調整する。CHIF21におけるプロトコル変換および速度制御の完了後、データはDCC22によるデータ転送制御を受け、CAdp24に転送され、CAdp23によりキャッシュメモリ7内に格納される。この時、CPU1から送られてきたデータが、論理アドレスの場合は、読み出しと同様にアドレステーブルによりアドレス変換を行い、物理アドレスに変換する。また、CPU1から送られてきたデータが新データの場合は、キャッシュメモリ7に格納したアドレスをアドレステーブル内のキャッシュアドレス31に登録する。この時、書き込む新データをキャッシュメモリ7内に保持するときは、論理アドレス27のキャッシュフラグ32をオン（1）とし、保持しない場合はキャッシュフラグ32をオフ（0）とする。なお、キャッシュメモリ7内に保持されている新データに対し、さらに書き込み要求がCPU1から発行された場合は、キャッシュメモリ7内に保持されている新データを書き替える。

【0037】キャッシュメモリ7に格納された新データは、この新データにより新しくパリティを更新し（以下更新されたパリティを新パリティとする）、以下のように論理グループ10内のドライブ12へ新データを格納し、FMEM41に新パリティを格納する。まず、すでにドライブ12内に書き込まれているデータを新しいデータに書き換える更新の場合についてのフローを図12を用いて示す。本実施例では論理グループ10において、RAIDのレベル4の制御を行い、パリティはパリティ専用のFMEM41に格納する。

【0038】本実施例の書き込み処理方法を図16を用いて説明する。MP20はCPU1が指定した論理アドレスからアドレステーブルを参照し、データが格納されているドライブ12（DDrive No. 29で指定される）とそのドライブ12内の物理的なアドレスであるSCSI内Addr30とパリティが格納されているFMEMチップ40のアドレスであるFMEMアドレス34'と、このFMEMチップ0内の物理アドレスであるFMEM内Addr35'を認識する。図16に示すようにCPU1からSD#1のドライブ12のData#1に対し、New Data#1に更新する書き込み要求が発行された場合、MP20はアドレステーブルにより更新されるデータ（旧データ）であるData#1および更新されるパリティ（旧パリティ）であるParity#1の物理アドレスを認識した後、旧データの格納されているドライブ12と旧パリティの格納されているFMEMチップ40に対し読み出しを行なう

（図16、図12の（1））。この時の旧データの読み出し方法は先に説明した読み出し処理におけるドライブ12からキャッシュメモリ7への読み出しと同じであ

る。FMEM41からの旧パリティの読み出しは以下のように行う。MP20は図17に示すアドレステーブルを参照し、更新される旧パリティのPアドレス33に対応する、旧パリティが格納されている物理アドレス（FMEMアドレス34'と、このFMEMアドレス34'のFMEMチップ40内のFMEM内Addr35'）を認識する。MP20がこのように更新される旧パリティのFMEM41内の物理アドレスを認識した後は、FMEM41内のFMEMC42に対し、当該旧パリティの読み出しコマンドとこの物理アドレスを送る。読み出しコマンドと物理アドレスを受け取ったFMEMC42では、FMEMアドレス34'に対するFMEMチップ40をイネーブルにし、FMEM内Addr35'を当該FMEMチップ40にセットし、当該旧パリティを読み出し、FMEMC42内のバッファに格納する。当該旧パリティを読み出したFMEMC42では、MP20に対し当該旧パリティの格納されている当該FMEMチップ40からの読み出しが完了したことを報告する。この報告を受け取ったMP20はFMEMC42に対し、当該旧パリティをキャッシュメモリ7に転送するように指示し、この指示を受け取ったFMEMC42は当該旧パリティをキャッシュメモリ7に転送する。ただ、書き込み時の読み出しでは、ADC2のMP20が発行した読み出し要求のため、読み出したデータはCPU1へは転送せず、キャッシュメモリ7に転送するのみである。この様に読み出した旧データ、旧パリティと書き込む新データとで排他的論理和を行ない更新後の新パリティであるNew Parity#1を作成しキャッシュメモリ7に格納する（図16、図12の

(2)）。新パリティ(New Parity#1)のキャッシュメモリ7への格納完了後、MP20は新データ(New Data#1)をSD#1のドライブ12のData#1のアドレスに書き込む（図16、図12の(3)）。なお、この新データの書き込みはMP20の管理の下で非同期に行なわれるようにしてもよい。新パリティ(New Parity#1)はキャッシュメモリ7にそのまま格納しておく。この時、図17に示すアドレステーブルに対し論理アドレスがData#1のエントリにNew Data#1を登録し、キャッシュメモリ7に保持しておく場合はキャッシュアドレス31にキャッシュ内のアドレスを登録し、キャッシュフラグ32をオンとする。また、パリティに関してはキャッシュメモリ7に保持したままのため、Pキャッシュアドレス36にキャッシュアドレスを登録し、Pキャッシュフラグ37をオンとする。なお、この様にアドレステーブルでPキャッシュフラグ37がオンとなっているパリティは、更新済みのパリティとなり、FMEM41内に格納されているパリティは無効とされる。本実施例では図5に示すように、CPU1からの新データを不揮発化されたキャッシュメモリ7内の領域に格納し、新パリティ

の作成が完了しキャッシュメモリ7に格納した時点で、MP20は書き込み処理を終了したとCPU1に報告する。なお、従来方法では図5に示したように、新パリティをドライブの1回転後に書き込み、MP20が書き込み処理を終了したとCPU1に報告している。新パリティのFMEM41への書き込みはMP20の管理の下で非同期に行なわれるため、ユーザからは見えない。また、新データのドライブ12への書き込みをMP20の管理の下に非同期に行なう場合は、同様にしてユーザからは見えない。以後CPU1からData#10、Data#8に対する書き込み処理が発行されれば、上記と同様に処理し、各新パリティをキャッシュメモリ7に格納していく。

【0039】キャッシュメモリ7に溜められた新パリティは、予めユーザが設定した設定値以上の新パリティがキャッシュメモリ7に溜った場合か、または、ユーザからの読み出し/書き込み要求の発行されていないタイミングが生じた場合にパリティ格納用のFMEM41にまとめて書き込む（図16、図12の(4)）。このように新パリティをパリティ格納用のFMEM41にまとめて書き込む場合は、シーケンシャルに書き込まれる。また、この様に新パリティをパリティ格納用のFMEM41に書き込む際に、アドレステーブルのFMEMアドレス34'、FMEM内Addr35'に実際に新パリティを書き込んだアドレスを登録する。従来のディスクアレイは全てドライブで構成され、データもパリティもドライブに格納されていた。また、新データは旧データが格納されていたアドレスに書き戻されるため、更新による書き込み時にアドレステーブルの変更を必要とせず、制御は簡単であった。しかし、本実施例では、新パリティは書き込み要求が発行された順にシーケンシャルにまとめて書き込まれるため、パリティ格納用のFMEM41内では、新パリティが格納されるアドレス（FMEMアドレス34とFMEM内Addr35）は原則として、その新パリティに対する旧パリティが格納されていたアドレスとは同一にせず、異なるものになっている。

【0040】また、キャッシュメモリ7内に溜められた新パリティをシーケンシャルにまとめて書き込む最中にCPU1より読み出し要求が発行された場合、読み出し処理にはパリティは関与しないため、先に説明したように通常の読み出し処理を行なう。一方、キャッシュメモリ7内に溜められた新パリティをシーケンシャルにまとめて書き込む最中にCPU1より書き込み要求が発行された場合は、通常の書き込み処理と同様に旧データを読み出し、旧データの読み出し後新データを書き込む。この時、キャッシュメモリ7には新データと旧データを保持し、キャッシュメモリ7内に溜められた新パリティをシーケンシャルにまとめて書き込む処理が終了次第、当該書き込まれた新パリティとで新たにパリティを更新する。

【0041】本実施例では新パリティをまとめ、シーケンシャルに書き込むが、図19、20に示すように、前に書き込まれているパリティが有効な場合、その上に新パリティを書き込んで消すわけには行かない。本実施例では、アドレステーブルでPキャッシュフラグ37がオンとなっている旧パリティは、更新済みのパリティとなり、パリティ格納用のFMEM41内では無効となっているが、Pキャッシュフラグ37がオフとなっているパリティはFMEM41内においてまだ有効なパリティである。この有効なパリティが消されると、ドライブ障害が発生した場合、障害ドライブ内のこのパリティの作成に関与したデータの回復が不可能となる。

【0042】そこで、新パリティをシーケンシャルにまとめ書きする時のパリティ格納用FMEM41内の有効データの扱い方について以下に説明する。図19に示すようにパリティ書き込み前のFMEMチップ40においてパリティのP1、P2、P3はデータの書き込み要求に伴い更新された無効パリティであり、P8、P9はデータに対する書き込み要求が発行されていないため、更新されていない有効パリティである。書き込み要求1、2、3の順に書き込み要求が発行され、これによりP2、P3、P1の順に旧パリティが更新され更新済みの新パリティとしてP'2、P'3、P'1の順にキャッシュメモリ7に格納されているとする。これらの新パリティをパリティ格納用のFMEM41にシーケンシャルにまとめ書きする場合は、MP20は書き込む新パリティの量を調べ、この量に相当するFMEMチップ40内の無効パリティ(P1、P2、P3)が書き込まれているアドレスを一度に消去する(フラッシュメモリの書き込みにおける消去方法についてはHN28F1600シリーズのデータシート(ADJ-203-045

(a)(z))に記載されている)。このように、新パリティをシーケンシャルにまとめ書きするFMEMチップ40内の消去が完了すると、旧パリティP1のアドレスに新パリティP'2を書き込み、有効パリティであるP8はそのまま残し、旧パリティP2の位置に新パリティP'3を書き込み、旧パリティP3の位置に新パリティP'1を順に書き込んでいく。すなわち、旧パリティの一連の位置に新パリティをその書き込み要求の発行順に順次書き込むのである。以上のように本発明ではシーケンシャルにまとめ書きする際に、有効データをそのまま残し、飛ばして新パリティを書き込んでいく。

【0043】また、別の方法としては図20に示すように有効パリティであるP8、P9をMP20の指示により、FMEM41に対し擬似的な読み出し要求を発行し、この擬似的な読み出し要求によりキャッシュメモリ7に有効パリティを読み出し、この有効パリティのキャッシュメモリ7への読み出しによりMP20はアドレステーブルのPキャッシュアドレス36をセットしPキャッシュフラグ37をオンとすることで、更新する新パ

ティとみなす。MP20はキャッシュメモリ7に溜められている新パリティと読み出された有効パリティの量を調べ、この量に相当するFMEMチップ40内の領域を一度に消去する。このように、FMEMチップ40内の消去が完了すると、FMEM41から読み出した有効パリティは新パリティとして、他の新パリティと一緒にシーケンシャルにまとめて書き込む。すなわち、図20に示すように、更新する新パリティP'2、P'3、P'1と更新する新パリティとみなしたパリティP8、P9からなる更新パリティ群の書き込み順をP'2、P'3、P'1、P8、P9とし、書き込み前のパリティ群P1、P8、P2、P3、P9の一連のアドレスへ上記書き込み順にしたがって更新パリティ群を順次書き込む。書き込み結果は図20のパリティ書き込み後のトラックに示すようになる。フラッシュメモリではデータをフラッシュメモリに書き込む場合、まず書き込むアドレスに格納されているデータを消去し、この消去が完了した後に実際にデータを書き込む。フラッシュメモリでは1セクタ(フラッシュメモリをアクセスする場合、ディスクにアクセスするときのアドレスと同じフォーマットのアドレスでアクセスする)を消去しても、複数セクタを一度に消去しても消去時間は同じである。又、書き込み時間の大半はこの消去時間が占めており、実際にフラッシュメモリに書き込む時間は消去時間と比較した場合無視できるほど小さい。本実施例の特徴であるシーケンシャルにまとめて書き込みを行うことにより、消去が一回で済むため、新パリティのまとめる数が多いほどオーバーヘッドが小さく出来る。

【0044】本実施例では、FMEM41を構成する全FMEMチップ40に対し、平均に書き込みが行われるようにFMEMチップ40への書き込み回数をカウントする。フラッシュメモリはドライブと比較してランダムにアクセスしても短時間で処理できる利点があるが、その反面書き込み回数に制限がある。このため、あるアドレスに集中して書き込みが行われると、フラッシュメモリの一部が書き込み回数の限界に達してしまい、以後書き込めなくなってしまう。本実施例では、FMEM41に対しシーケンシャルにまとめて新パリティを書き込むため、FMEM41へ新パリティをシーケンシャルにまとめて書き込む際に、規則的に書き込むことで全FMEMチップ40に対し書き込み回数を平均化する。具体的には、図18に示すようにFMEM41がFMEM#1、2、3、4、...、nのn個のFMEMチップ40で構成され、アドレスが0000からffffとする。MP20は新パリティのシーケンシャルなまとめ書きを行うことを認識すると、前回のまとめ書きを行った際に、新パリティを格納した最終のアドレスを調べる。例えば、前回のまとめ書きにおいて、FMEM41内で0000からaaaaまで新パリティを書き込んだとすると、アドレスaaaaをMP20は記憶してお

く。そして、次のシーケンシャルなまとめ書き時にはこの記憶しておいたアドレス (aaaa) を調べる。このように、MP20が前回の最終アドレスを記憶すると、MP20は次の新パリティのシーケンシャルなまとめ書きはアドレスaaaaの次から行うように判断する。MP20がこのようにして新パリティのシーケンシャルなまとめ書きを行う先頭アドレスを認識すると、次にFMEM41への書き込み回数の限界に達しているかどうかを判定する。

【0045】図22の書き込み回数判定フローチャートに示すように、新パリティをシーケンシャルにまとめ書きする先頭アドレスが0000かどうかを判定する(51)。0000でなければ判定フローは終了し(52)、0000であればMP20は書き込み回数カウンタのカウント値に1をたす(53)。すなわち、先頭アドレスに来る度にカウンタの値を増加する。次にMP20はこの1をたしたカウンタ値が、予めセットされているFMEMチップ40の書き込み回数の限界値かどうかを判定する(54)。FMEMチップ40の書き込み回数の限界値は、初期設定において予めMP20に対しユーザが設定する。この判定結果において、FMEMチップ40の書き込み回数の限界値を越えていなければ判定フローを終了し(52)、越えている場合はFMEMチップ40の交換を通報する(55)。以上のように本実施例ではFMEM41に対する新パリティのシーケンシャルなまとめ書きは、アドレスの低い方から高い方へ方向に行われる。これにより、FMEM41の全FMEMチップ40では、書き込み回数は平均化されることになる。本実施例ではこのようにしてFMEM41内においてFMEMチップ40への書き込み回数を平均化する。

【0046】一方ドライブ12にすでに格納されているデータに新しいデータ追加する新規書き込みの場合は、MP20はアドレステーブルにおいて空き領域を探す。空き領域には2種類ある。まず一つはまったく使用されていない未使用領域である。この様にまったく使用されていない領域では、アドレステーブルにおいて論理アドレス27の項に論理アドレスは登録されていない。このため、MP20はアドレステーブルにおいて論理アドレス27の項に論理アドレスが登録されていない領域を探すことで、未使用領域を見つけられる。もう一つの空き領域は、以前その領域は使用されていたが(データが書き込まれていた)、ユーザがそのデータが必要でなくなったため削除した削除領域である。削除領域は、アドレステーブルにおいて論理アドレス27の項に論理アドレスが登録されているが無効フラグ28をオン(1)としている。このため、MP20はアドレステーブルにおいて無効フラグ28がオンになっている領域を探すことで、削除領域を見つけられる。MP20が新規書き込みを行なう空き領域を決定する場合、まず、未使用領域を探

す。もし、未使用領域が無い場合は削除領域を新規書き込み先に使用する。これは、未使用領域はパリティの作成に関与していない(全て0で構成されているとした)ため、新規書き込みの際のパリティの更新は、新規書き込みする新データと更新される旧パリティとの排他的論理和のみで行なえるが、削除領域のデータはユーザにとっては意味が無いデータとなっているが、パリティの作成には関与しているため、新規書き込みの際に旧データと同じように読み出して、旧パリティと新規書き込みデータとの間で排他的論理和をとり新パリティを作成しなければならない。このため、未使用領域に新規書き込みを行なうのと、削除領域に新規書き込みを行なうのでは、削除領域から削除されたデータを読み出す処理が入らない分、未使用領域に新規書き込みを行なう方が早く処理できるためである。以上述べたようにMP20が空き領域を探し、空き領域の認識が完了した後、この空き領域に新規書き込みデータの書き込みを行ない、更新と同様にアドレステーブルに論理アドレス27を登録し、削除領域に新規書き込みを行なった場合は無効フラグ28をオフとする。以上述べたように、新規書き込みと更新では、新データの書き込み先が異なるのみで処理自体は同じである。

【0047】(障害回復処理)次にドライブ12に障害が発生した場合の、障害ドライブ12内のデータを回復する方法を説明する。図16に示すようにSD#1のドライブ12のData#1とSD#2のドライブ12のData#2とSD#3のドライブ12のData#3とSD#4のドライブ12のData#4からFMEM#1のFMEMチップ40のParity#1が作成されている。同様にData#5、6、7、8からParity#2、Data#9、10、11、12からParity#3が作成されている。SD#1、2、3、4のドライブ12の中でどれか1台のドライブ12に障害が発生した場合、残りのドライブ12内のデータとFMEM41内のパリティから、障害ドライブ12内のデータを回復する。本実施例では、パリティはパリティを格納するFMEM41内においてランダムなアドレスに格納されている。例えば、図16においてSD#1のドライブ12に障害が発生し、このSD#1のドライブ12内のData#1に読み出し要求が発生したとする。まず、MP20はアドレステーブルから読み出し要求が発行されたData#1は障害が発生したドライブ内のデータだと認識した場合は、SD#2、3、4のドライブ12からData#2、3、4をそれぞれ読み出し、MP20はアドレステーブルによりこれらのデータに対応するパリティを、アドレステーブルにより探す。MP20がアドレステーブルにより当該パリティのFMEM41内のアドレスを認識した後はFMEM41から当該パリティであるParity#1を読み出し、上記データと共にパリティ生成回路(PG)25に送り、排他的論

理和を行うことでData #1を復元する。同様にData #5、9も復元する。フラッシュメモリはドライブのようにディスクを回転させたり、ヘッドをシークさせるような機械的な動作を必要としない。また、フラッシュメモリでは先に述べたように書き込み時は消去を必要とするが、読み出し時にはDRAMのような半導体メモリからの読み出しと同じように短時間で読み出せる。このようにフラッシュメモリにおけるランダムな読み出しはドライブと比較して無視できるほど短時間で処理できる。このため、本実施例のようにデータの回復毎にアドレステーブルを調査し、アドレスを認識し、FMEM41から読み出しても、ドライブからのデータの読み出しと比較し短時間で処理できる。この様に復元したデータは、障害ドライブ12を正常なドライブ12に交換した後、この正常なドライブ12に書き込むことで回復処理を行なう。また、ドライブ12の障害時に備え予め予備の正常なドライブ12を用意してある場合は、この予備の正常なドライブ12に復元したデータを書き込みことで回復処理を行なう。

【0048】以上の説明では更新後の新パリティを格納するキャッシュメモリ7は不揮発な半導体メモリとした。しかし、パリティはデータとは異なり停電等によってキャッシュメモリ7から消失しても、新たに作り直すことが可能なため、この、新たに作成する手間を許容できるなら、キャッシュメモリ7内で旧パリティを格納する領域を揮発な半導体メモリにすることも可能である。以上の説明では、更新後の新パリティをキャッシュメモリ7に格納したが、キャッシュメモリ7ではなく専用のメモリを用意することも可能である。従来のレベル4、5では新パリティをドライブに格納しているため、書き込み処理を行なうたびに新パリティの書き込みを行なっていたため、新パリティを書き込む毎に回転の回転待ちを必要としたが、本実施例ではシーケンシャルなまとめ書きを行なう際のFMEMチップ40の一括消去(約10ms)と書き込みを行う時間のみである。

【0049】(実施例5)本実施例では実施例4で示したように、1つのパリティ格納用のFMEMチップにシーケンシャルにまとめ書きするのではなく、複数のパリティ格納用のFMEMチップ40に対し新パリティを並列に書き込む方法を示す。本実施例でも実施例4と同じ処理により、データの書き込みに伴いパリティを更新し、更新した新パリティはキャッシュメモリ7に保持される。図21に示すように書き込み要求1、2、3によりData #1、#9、#8がそれぞれNew Data #1、#9、#8に更新され、このデータの更新により更新された新パリティとしてNew Parity #1、#3、#2がキャッシュメモリ7に保持されている(図21の(1)(2)(3))。実施例1と同様に予めユーザの設定値以上の新パリティがキャッシュメモリ7に溜った場合か、または、ユーザからの読み出し

／書き込み要求の発行されていないタイミングで、複数のパリティ格納用のFMEM41であるFMEM #1、FMEM #2に並列にまとめて書き込む(図21の(5))。並列にまとめて書き込む単位としては、レベル3のようにバイト単位と、レベル4、5のようにパリティ単位がある。この時、各パリティ格納用のFMEM41に対する書き込み方法は、実施例1のパリティ格納用のFMEM41への書き込み方法と同じである。また、本実施例の変形として、並列に書き込む新パリティによりパリティを作成し、FMEM #3のパリティ格納用のFMEM41に書き込む。この様にパリティのパリティを作成することにより、パリティ格納用のFMEM41の障害時に新たにパリティを作成する際に、データを読み出す必要が無く、その間のデータへのアクセスを受け付けることが可能となる。

【0050】

【発明の効果】従来はデータの書き込みによるパリティの更新ごとにドライブに書き込むため、この更新された新パリティの書き込みに1回転の回転待ちが必要であった。本発明によれば、更新した新パリティをキャッシュメモリに貯蔵しておき、後に、この様に貯蔵された新パリティをまとめてシーケンシャルにパリティ格納用のドライブに書き込んでいるため、シーケンシャルにまとめ書きする際の、最初に0.5回転の回転待ちが必要になるのみであり、従来問題となっていた書き込み時の処理時間を減少させることが可能となる。また、更新した新パリティをキャッシュメモリに貯蔵しておき、後に、この様に貯蔵された新パリティをまとめてシーケンシャルにパリティ格納用のFMEMに書き込んでいるため、シーケンシャルにまとめ書きする際の、FMEMの消去時間と書き込み時間のオーバーヘッドが必要になるのみであり、書き込み時の処理時間を減少させることが可能となる。また、パリティを格納するFMEMの書き込み回数の平均化が可能になる。

【図面の簡単な説明】

【図1】実施例1のハードウェア構成を示す図である。

【図2】図1のチャンネルバスディレクタと1クラスタの内部構造を示した図である。

【図3】実施例1の書き込み処理時におけるデータ移動を説明する図である。

【図4】アドレス変換テーブルを説明する図である。

【図5】書き込み処理のタイミングチャートを示す図である。

【図6】磁気ディスク装置内の領域分割を説明する図である。

【図7】新パリティのシーケンシャル書き込み方法を説明する図である。

【図8】新パリティの他のシーケンシャル書き込み方法を説明する図である。

【図9】実施例2の書き込み処理時におけるデータ移動

を説明する図である。

【図10】RAIDのレベル4、5における更新処理を説明する図である。

【図11】磁気ディスク装置内のアドレスを説明する図である。

【図12】データおよびパリティの書き込み処理のフローチャートを示す図である。

【図13】実施例3における磁気ディスク装置内の領域分割を説明する図である。

【図14】実施例4のハードウェア構成を示す図である。

【図15】図14のチャンネルバスディレクタと1クラスターの内部構造を示した図である。

【図16】実施例4の書き込み処理時におけるデータ移動を説明する図である。

【図17】実施例4におけるアドレス変換テーブルを説明する図である。

【図18】FMEM内のアドレスを説明する図である。

【図19】実施例4における新パリティのシーケンシャル書き込み方法を説明する図である。

【図20】実施例4における新パリティの他のシーケンシャル書き込み方法を説明する図である。

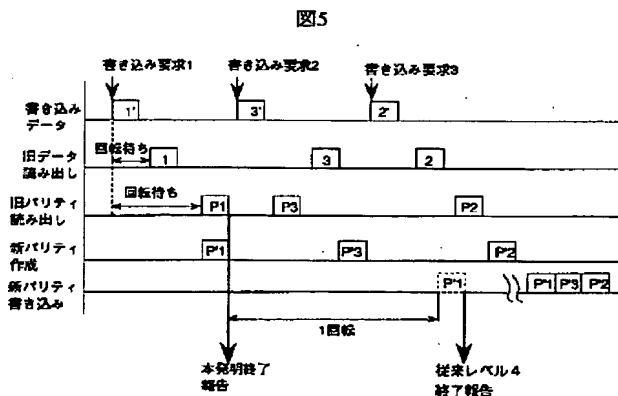
【図21】実施例5の書き込み処理時におけるデータ移動を説明する図である。

【図22】FMEMの書き込み回数を判定するフローチャートを示す図である。

【符号の説明】

- 1 CPU
- 2 ディスクアレイコントローラ (ADC)
- 3 ディスクアレイユニット (ADU)
- 4 外部インターフェースバス
- 5 チャンネルバスディレクタ
- 6 チャンネルバス
- 7 キャッシュメモリ

【図5】



- 8 ドライブバス
- 9 ディスクアレイユニットバス
- 10 論理グループ
- 12 ドライブ
- 13 クラスタ
- 14 ドライブ側キャッシュアダプタ (C Adp)
- 15 インターフェースアダプタ
- 16 チャンネルバススイッチ
- 17 制御信号線
- 18 データ線
- 19 バス
- 20 マイクロプロセッサ (MP)
- 21 チャンネルインターフェース回路 (CH IF)
- 22 データ制御回路 (DCC)
- 23 チャンネル側キャッシュアダプタ (C Adp)
- 24 ドライブインターフェース回路 (Drive IF)
- 25 パリティ生成回路 (PG)
- 27 論理アドレス
- 28 無効フラグ
- 29 データドライブ番号 (DDrive No.)
- 30 SCSI内Addr
- 31 キャッシュアドレス
- 32 キャッシュフラグ
- 33 P論理アドレス
- 34 パリティドライブ番号 (PDrive No.)
- 34' FMEMアドレス
- 35' PSCSI内Addr
- 35' FMEM内Addr
- 36 Pキャッシュアドレス
- 37 Pキャッシュフラグ
- 40 フラッシュメモリチップ (FMEMチップ)
- 41 フラッシュメモリ (FMEM)
- 42 フラッシュメモリコントローラ (FMEMC)

【図17】

図 17

10	11	12	13	14	15	16	17	18	19
物理 アドレス	補助 Flag	Drive No.	SCSI 内Addr	キャッシュ アドレス	Cache Flag	P 物理 アドレス	FMEM アドレス	FMEM 内Addr	P キャッシュ アドレス
Data#1	0	SD#1	DADR1	—	0	Parity#1	FMEM#1	FADR5	C ADR21
Data#2	0	SD#2	DADR1	C ADR5	1				
Data#3	0	SD#3	DADR1	—	0				
Data#4	0	SD#4	DADR1	C ADR2	1				
Data#5	1	SD#1	DADR2	—	0	Parity#2	FMEM#1	FADR7	—
Data#6	0	SD#2	DADR2	C ADR8	1				
Data#7	0	SD#3	DADR2	C ADR1	1				
Data#8	0	SD#4	DADR2	—	0				

【图 2】

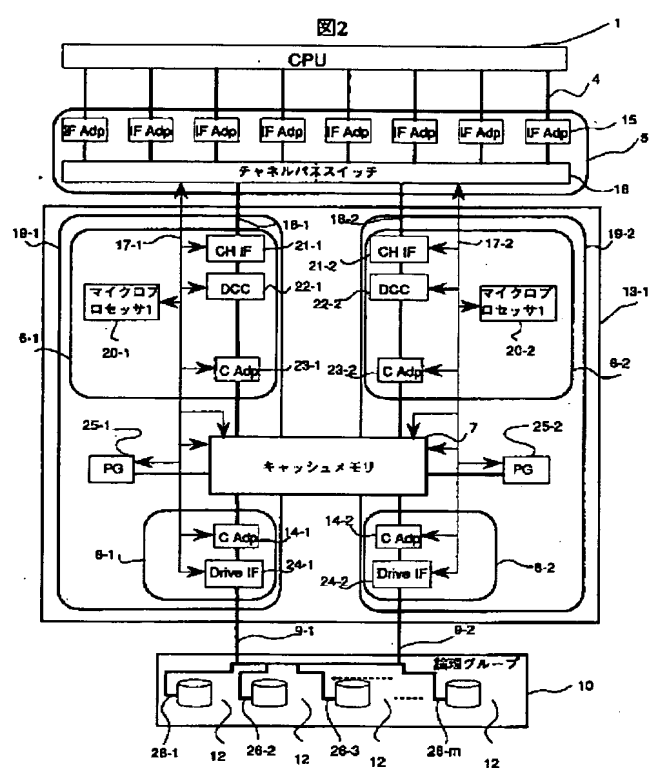
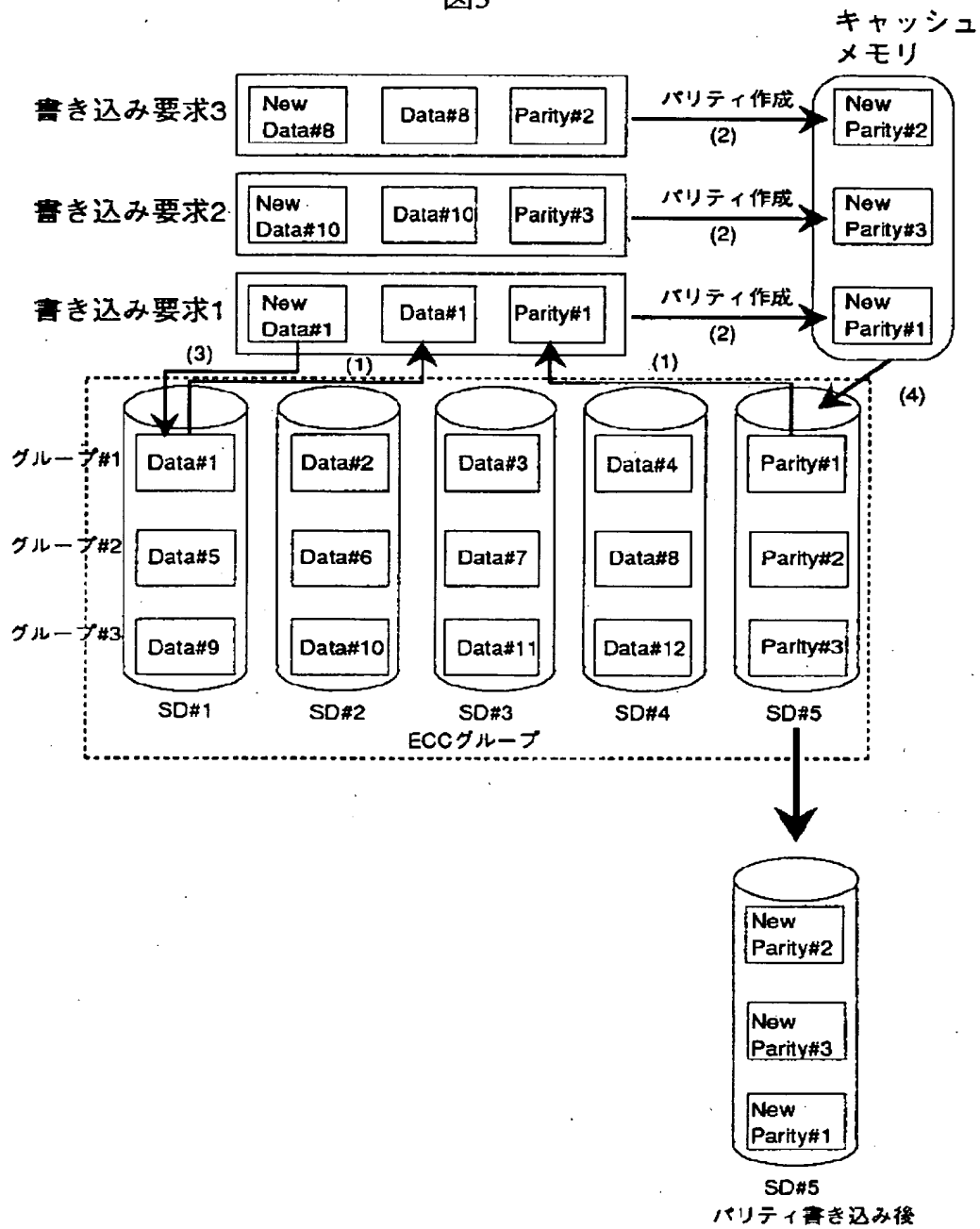


图4

27	28	29	30	31	32	33	34	35	36	37
論理 アドレス	無効 Flag	DDrive No.	SCSI 内Addr	キャッシュ アドレス	Cache Flag	P論理 アドレス	PDrive No.	PSCSI 内Addr	Pキャッシュ アドレス	PCache Flag
Data#1	0	SD#1	DADR1	—	0	Parity#1	SD#5	DADR5	C ADR21	1
Data#2	0	SD#2	DADR1	C ADR5	1					
Data#3	0	SD#3	DADR1	—	0					
Data#4	0	SD#4	DADR1	C ADR2	1					
Data#5	1	SD#1	DADR2	—	0	Parity#2	SD#5	DADR7	—	0
Data#6	0	SD#2	DADR2	C ADR6	1					
Data#7	0	SD#3	DADR2	C ADR1	1					
Data#8	0	SD#4	DADR2	—	0					
⋮	⋮	⋮	⋮	⋮	⋮					

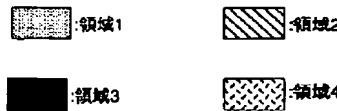
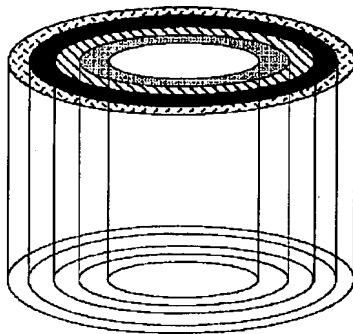
【図3】

図3



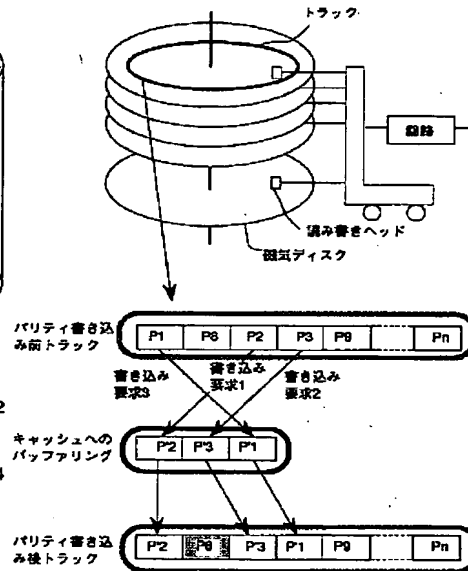
【図 6】

図6



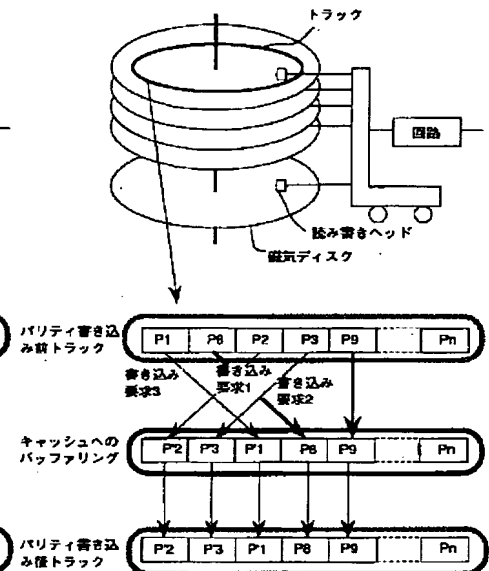
【図 7】

図7



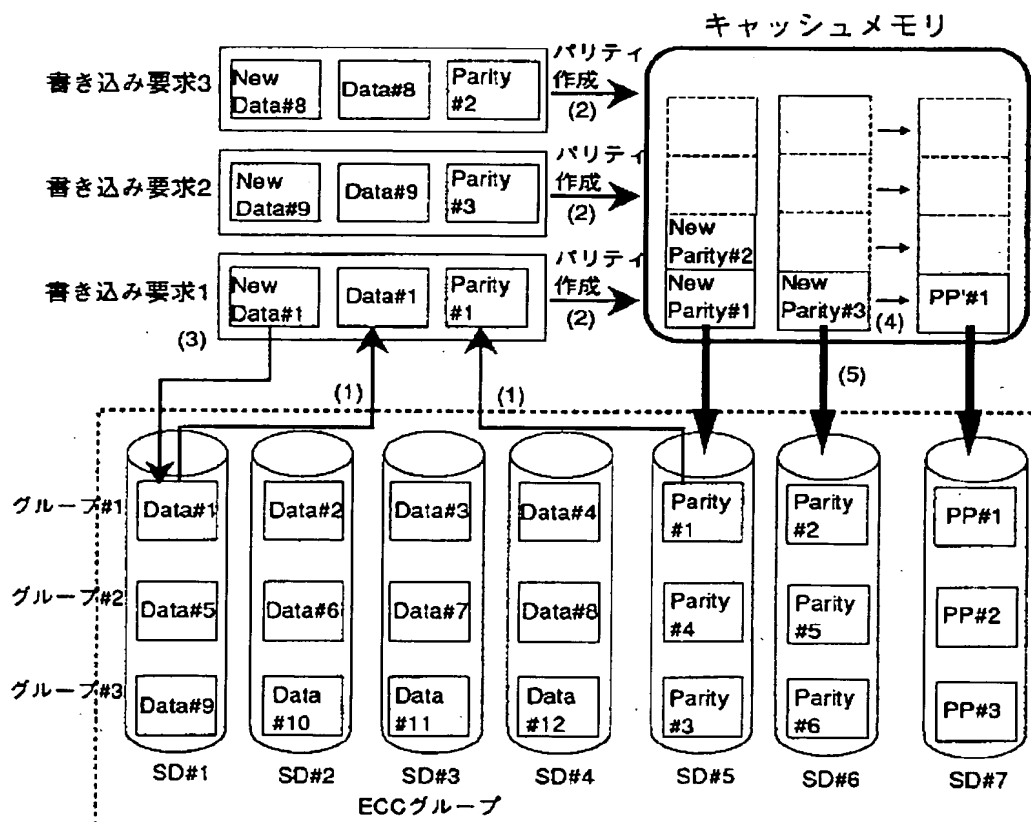
【図 8】

図8

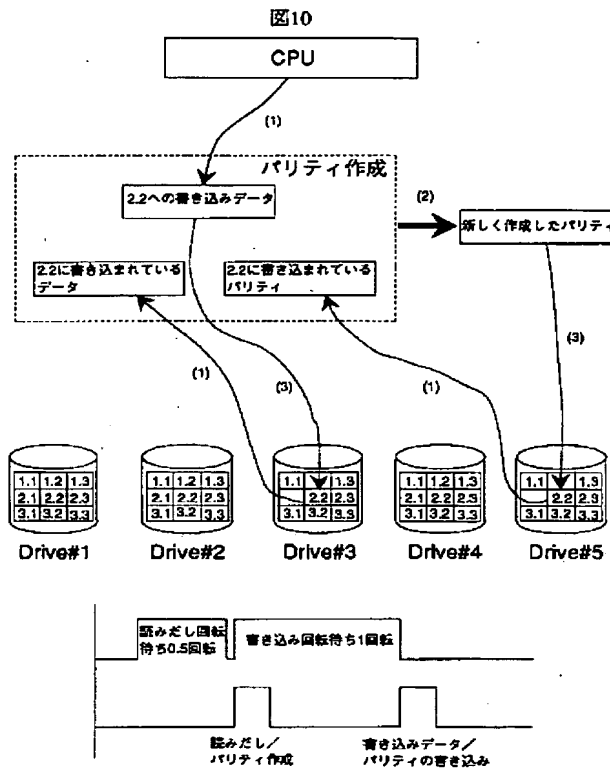


【図 9】

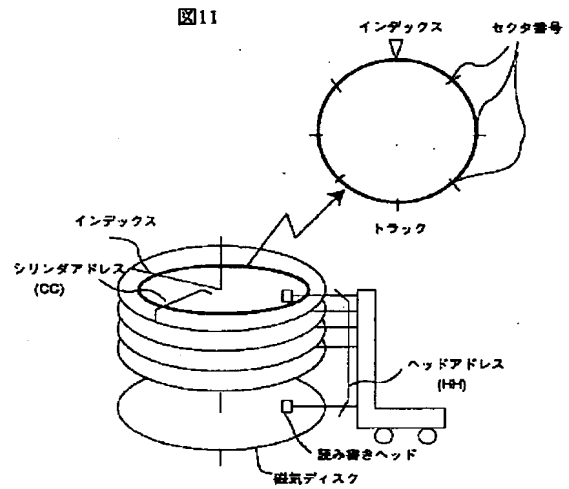
図9



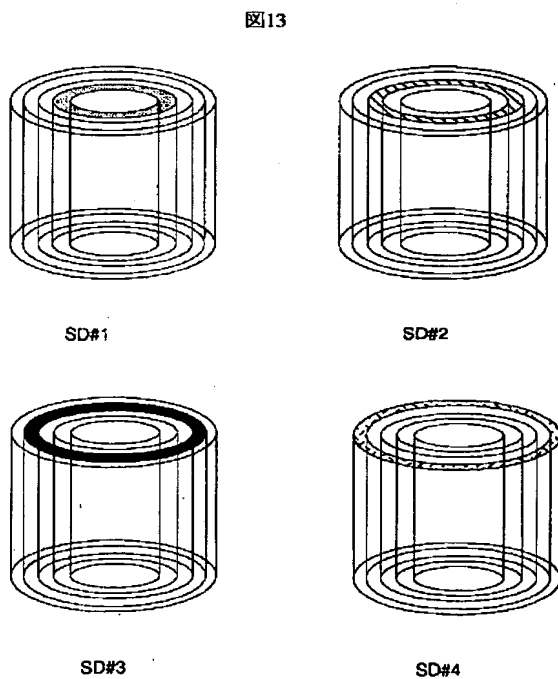
【図10】



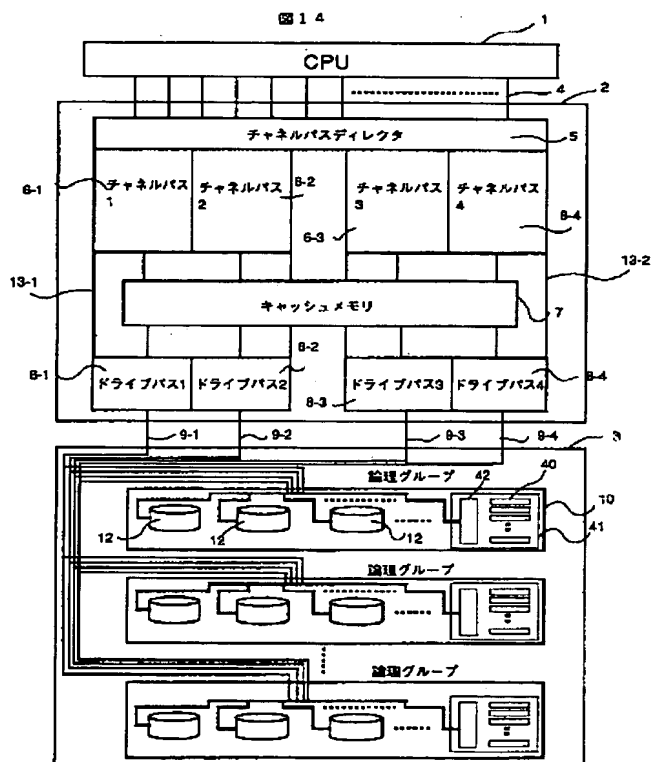
【図11】



【図13】

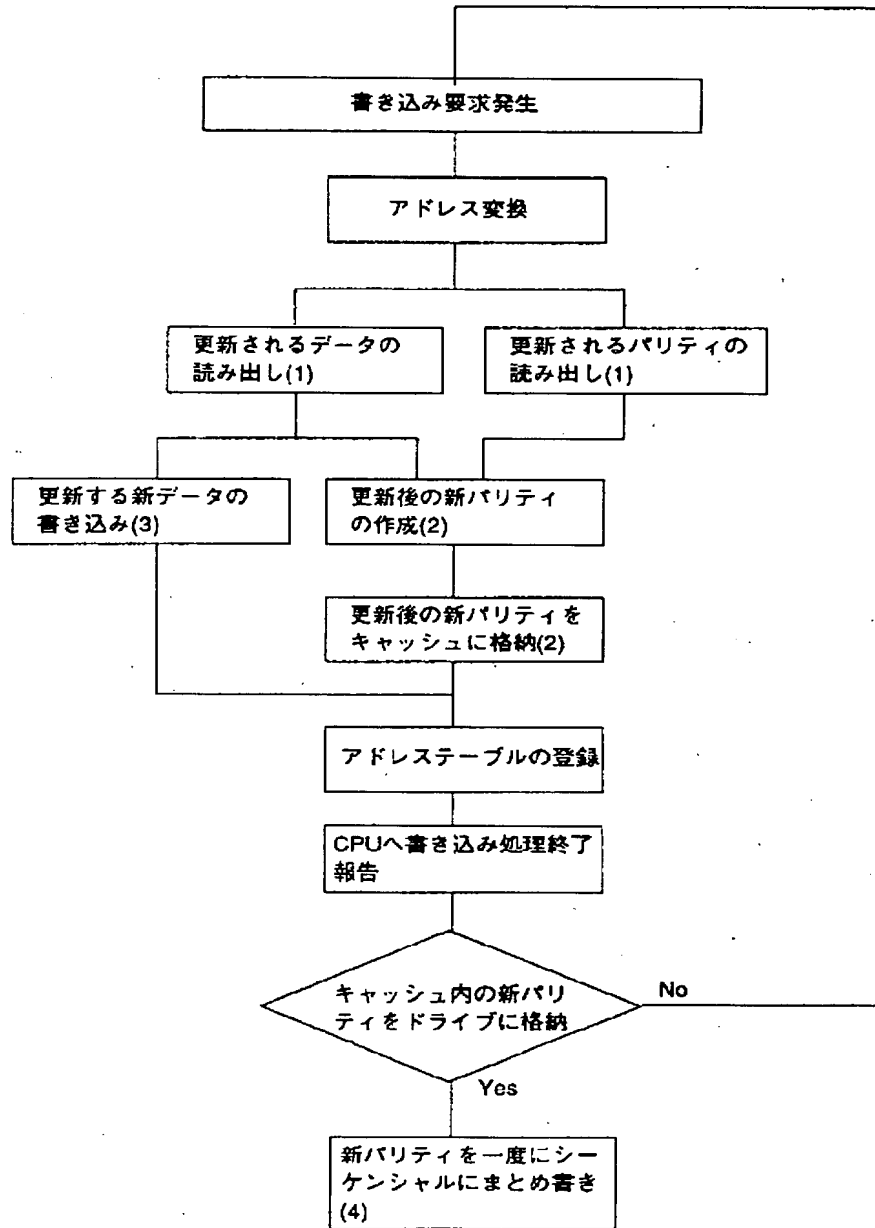


【図14】

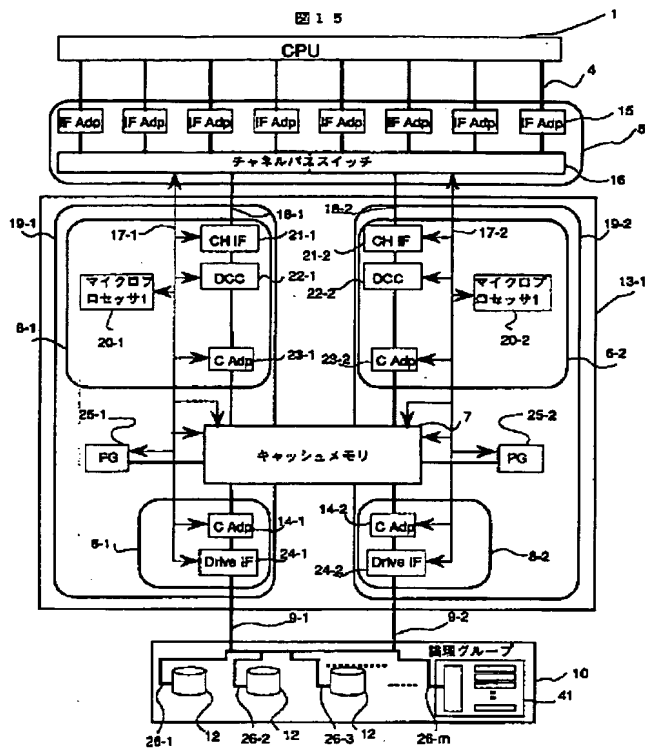


【図 1 2】

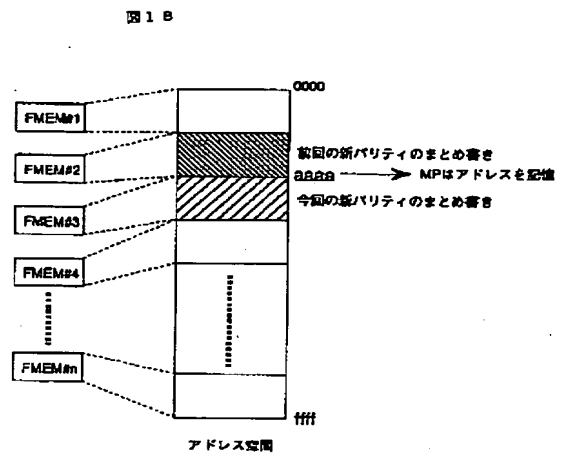
図12



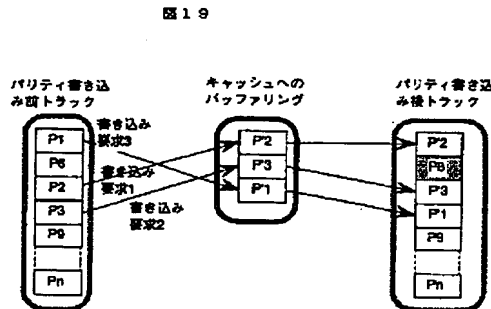
【図15】



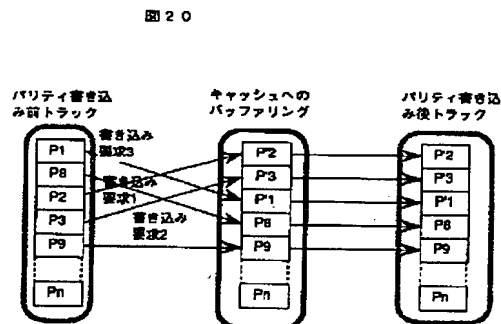
【図18】



【図19】

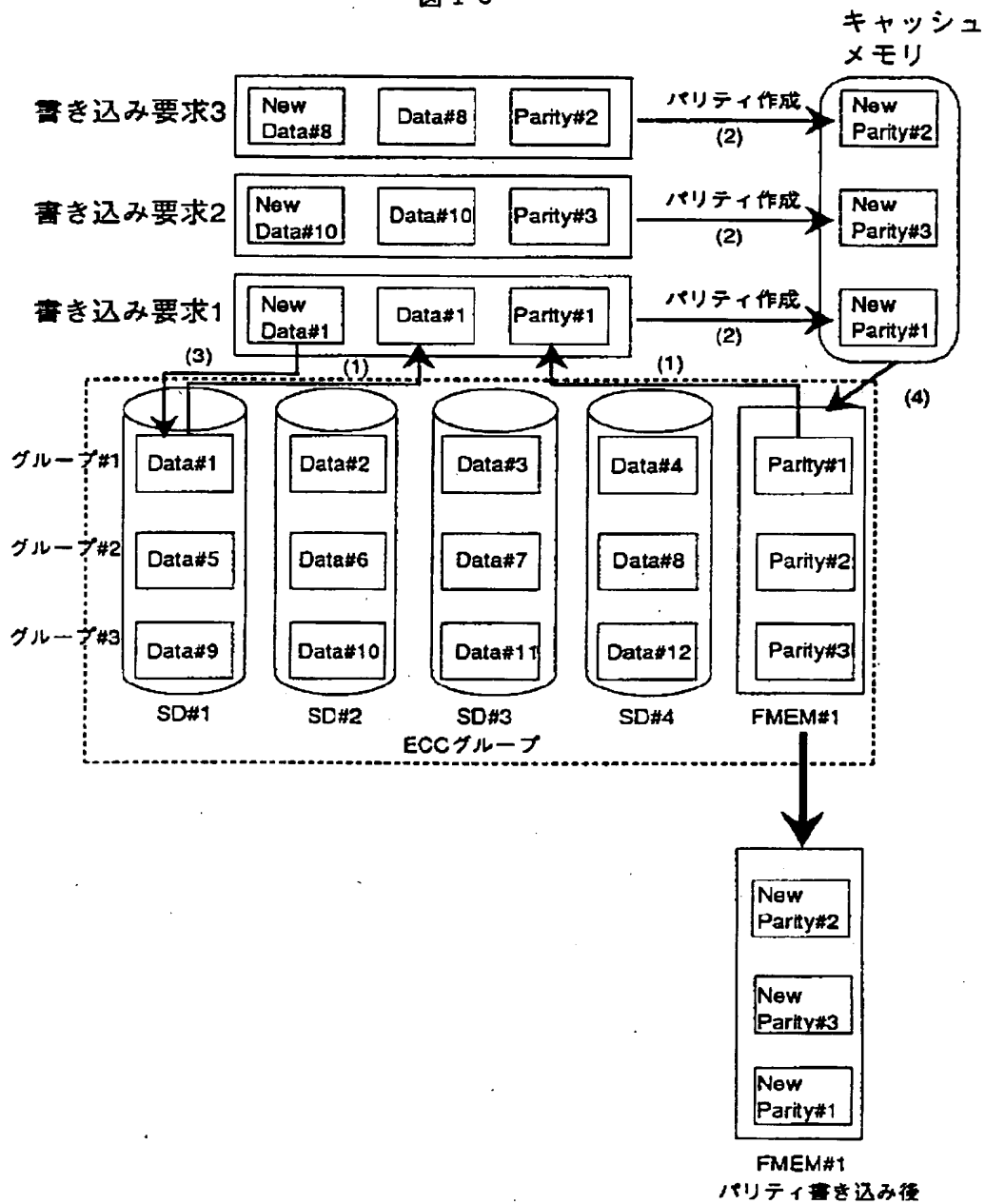


【図20】



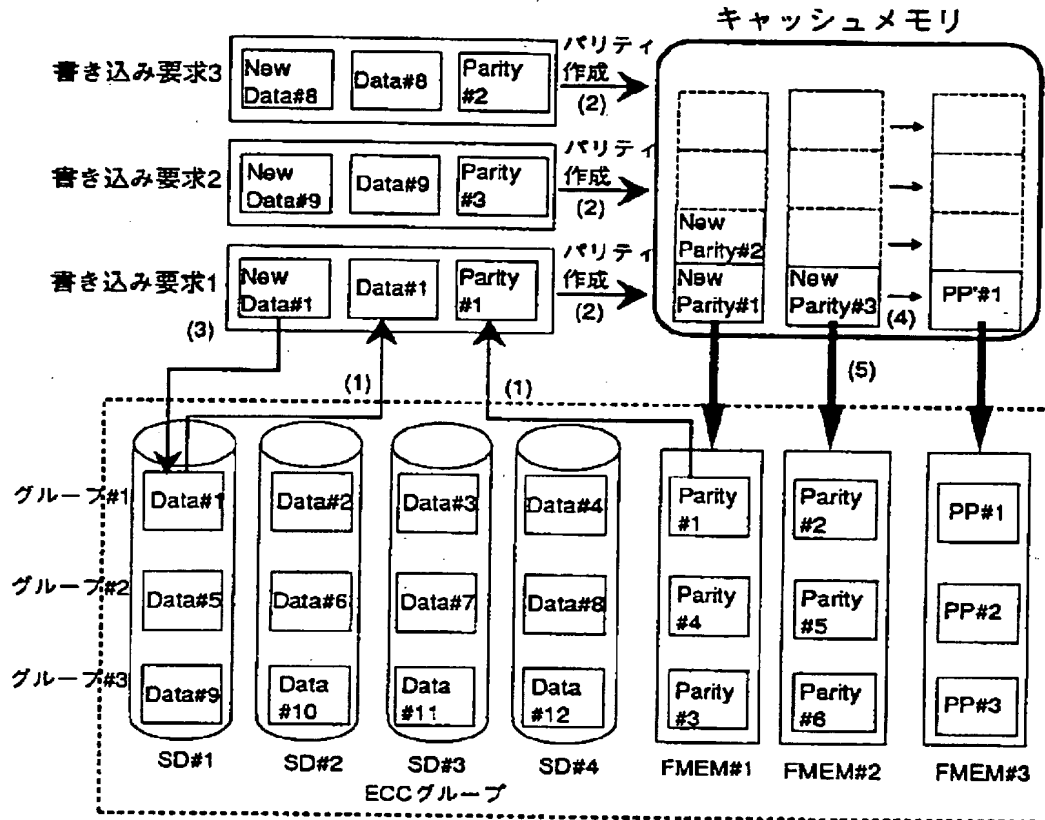
【図16】

図 16



【図 2 1】

図 2 1



【図 2 2】

図 2 2

